The Review of Financial Studies

Man versus Machine Learning Revisited

Yingguang Zhang
Guanghua School of Management, Peking University, China

Yandi Zhu
School of Finance, Central University of Finance and Economics, China

Juhani T. Linnainmaa
Dartmouth College, NBER, and Kepos Capital, USA

Binsbergen, Han, and Lopez-Lira (2023) predict analysts' forecast errors using a random forest model. A strategy that trades against this model's predictions earns a monthly alpha of 1.54% (t-value = 5.84). This estimate represents a large improvement over studies using classical statistical methods. We attribute the difference to a look-ahead bias. Removing the bias erases the alpha. Linear models yield as accurate forecasts and superior trading profits. Neither alternative machine learning models nor combinations thereof resurrect the predictability. We discuss the state of research into the term structure of analysts' forecasts and its causal relationship with returns. (*JEL* G12, G14)

Received: May 3, 2024; Editorial decision: April 17, 2025
Editor: Tarun Ramadorai
Authors have furnished an Internet Appendix, which is available on the Oxford University Press Web site next to the link to the final published paper online.

Are sell-side analysts' forecasts biased and, if so, can investors profit by trading against such biases? La Porta (1996) and So (2013) find that strategies that bet against analysts' long-term growth expectations and estimated biases are profitable. Neither strategy, however, has earned a reliable alpha since 2000.[1] These results could be interpreted as implying that analysts' biases might have correlated with (or even contributed to) mispricings but that markets have since turned more efficient. Indeed, after decades of research into analysts, the consensus has emerged that the link between biases in analyst expectations

We thank Tarun Ramadorai (editor) and four anonymous referees for their comments. Supplementary material can be found on The Review of Financial Studies web site. Send correspondence to Yandi Zhu, yandi.zhu@cufe.edu.cn.

[1] In the post-2000 data, a value-weighted version of the La Porta (1996) strategy earns an average return of 15 basis points per month (t-value = 0.34) when replicated using Chen and Zimmermann's (2022) data. The So (2013) strategy earns an average return of 33 basis points (t-value = 1.79) in the post-2000 data (see Section 1.4).

The Review of Financial Studies 38 (2025) 3768–3790

https://doi.org/10.1093/rfs/hhaf066 Advance Access publication September 2, 2025

and stock prices is economically and statistically weak (Kothari, So, and Verdi 2016).

Binsbergen, Han, and Lopez-Lira (2023) (BHL), whose study is the first to use machine learning techniques to model analyst forecasts, seemingly upend this literature. They train a random forest regression to serve as "a statistically optimal unbiased machine-learning benchmark." The difference between analysts' forecasts and this benchmark measures how overly optimistic or pessimistic analysts are in their forecasts. Their approach appears to prove extraordinarily fruitful. A strategy that buys (sells) stocks that receive too pessimistic (optimistic) earnings forecasts earns a monthly value-weighted Fama-French five-factor model alpha of 1.54% (t-value = 5.84).

This finding stands out. First, other studies that construct similar strategies find weaker results that grow weaker still in the more recent data. Second, even outside the analyst literature, their results dwarf all other anomalies by a significant margin in their sample. Third, Binsbergen, Han, and Lopez-Lira (2023) train the random forest regression using just 1 or 2 years of data at a time. If this technique outperforms So's (2013) protocol, it must be because the random forest regression uncovers meaningful interactions and nonlinearities from the data. However, with 75 features and a small sliver of training data, it seems, a priori, challenging for a random forest model to do so.[2]

If the Binsbergen, Han, and Lopez-Lira (2023) result is correct, it has far-reaching ramifications. As a first-order effect, any study that seeks to understand analysts' expectations (and the biases therein) must use or control for the Binsbergen, Han, and Lopez-Lira (2023) technology. Moreover, in the spirit of Harvey, Liu, and Zhu (2016), a t-value of 5.84 sets a high bar for what constitutes a statistically and economically meaningful alpha. In short, the Binsbergen, Han, and Lopez-Lira (2023) result would affect what would be publishable and, therefore, how researchers might choose to allocate their efforts.

In this study, we show that these inferences are premature. Table 1 explains our key point. In columns 1 and 2, we reproduce the estimates from Binsbergen, Han, and Lopez-Lira's (2023) Table 6. These estimates represent the profitability of a strategy that trades against predicted analyst forecast errors. (We detail this strategy later.) This is the strategy that earns an alpha of -1.54% per month. In columns 3 and 4, we replicate their result based on the details provided in the original paper and a search over plausible rules when the details are missing. We closely match the estimates, not only in terms of the alpha (-1.71% versus -1.54%) but also the factor loadings.

[2] The machine learning literature typically calls dependent and independent variables "targets" and "features." We use this nomenclature when referring to the random forest regression. The trade-off between the amount of training data and the quality of the model's predictions depends on model complexity. A model such as random forest requires more data to avoid overfitting to the training data and to make accurate predictions in the test data.

Table 1
Monthly Fama-French five-factor model alphas and factor loadings of portfolios sorted on analysts' conditional biases

			Replication			
	BHL Original		With Look-Ahead Bias		Without Look-Ahead Bias	
	Coef.	*t*-value	Coef.	*t*-value	Coef.	*t*-value
Intercept	−1.54	−5.84	−1.71	−6.07	−0.41	−1.54
MKT	0.38	5.28	0.43	5.71	0.44	5.94
SMB	0.61	5.17	0.66	5.14	0.75	5.69
HML	0.95	7.12	0.90	6.27	0.84	5.90
RMW	−0.68	−4.10	−0.75	−4.01	−0.72	−3.83
CMA	−0.53	−1.93	−0.30	−1.09	−0.18	−0.67

This table reports Fama-French five-factor model alphas and factor loadings for three long-short strategies. Each strategy is long stocks with the highest conditional earnings forecast biases and short stocks with the lowest biases. We measure the conditional biases in three steps, following Binsbergen, Han, and Lopez-Lira (2023). First, we measure the expected earnings benchmarks over the 1-quarter, 2-quarter, 3-quarter, 1-year, and 2-year horizons using random forest regressions. Second, we compute the differences between analysts' consensus earnings forecasts and these benchmarks at each horizon. Finally, we average these differences at the five horizons as the conditional bias. Columns 1 and 2 report the estimates from Table 6 in Binsbergen, Han, and Lopez-Lira (2023). Columns 3 and 4 show our replication of the result. Columns 5 and 6 remove the variable with the look-ahead bias from the random forest regression used to estimate the conditional biases. The sample period is January 1986 through December 2019.

Alas, our replication (and, by extension, the results in BHL) has a look-ahead bias. One of the 75 features is the company's actual earnings, described in BHL (p. 14) as "Realized earnings from the last period." However, this variable is from the "last period" from the perspective of the earnings being forecasted. This is a problem: when the target variable is, for example, the 2-years-ahead earnings, this variable is the firm's 1-year-ahead earnings, a variable unknown to econometricians and investors at time t. This look-ahead bias is important. In columns 5 and 6 we remove this bias by using realized earnings from the most recent period as of the forecasting date. We keep every other step of the analysis unchanged. This modification reduces the monthly alpha to −41 basis points and we can no longer reject the null hypothesis that this strategy's true alpha is zero (t-value = −1.54). Alphas computed from five alternative modern asset pricing models are close to zero with t-values ranging from −0.27 to 0.87.[3]

Binsbergen, Han, and Lopez-Lira (2023) present a number of results and conclusions related to the forecasting of earnings, the profitability of trading strategies and anomalies, and corporate decisions. Our critique above is largely from the viewpoint of the return predictability result. However, to guide future research, it is vital to catalogue how sensitive each result is to the look-ahead bias and, removing this bias, to what extent do the findings advance the prior literature. The study's results fall into four categories:

[3] These alternative models are the Fama-French-Carhart six-factor model, Hou, Xue, and Zhang's (2015) q-factor model, Stambaugh and Yuan's (2016) mispricing model, Daniel, Hirshleifer, and Sun's (2020) short- and long-horizon behavioral model, and Hou et al.'s (2021) q5-factor model.

1. **Accurate forecasting of earnings.** Binsbergen, Han, and Lopez-Lira's (2023) key result is about implementing a machine learning technique to construct a "statistically optimal and unbiased benchmark" for earnings. The look-ahead bias significantly enhances the accuracy of the machine learning model. What remains of the model does not significantly advance the literature. A linear model akin to that in Hughes, Liu, and Su (2008) predicts earnings as accurately as the random forest model.
2. **Profitable trading strategy.** Binsbergen, Han, and Lopez-Lira's (2023) main application for the objective benchmark is the strategy that trades against the estimated biases in analysts' forecasts. As noted above, this trading strategy does not attain statistical significance without the look-ahead bias under any modern factor models. Moreover, the corrected strategy is weaker than that in So (2013), which shares the idea of accurate forecasting of earnings.
3. **Overpricing and equity issuance.** The authors find that managers appear to issue more equity when forecasts are optimistic relative to their benchmark.[4] The reported numbers are not sensitive to the look-ahead bias. However, the reason is that it is not the estimated bias (= analyst forecast − predicted earnings) that predicts equity issuance but the predicted earnings. When we sort stocks into portfolios by the model's earnings forecasts, we find a larger spread in the issuance activity; and when we sort by the actual realized earnings, this spread becomes larger still. The bias component of the Binsbergen, Han, and Lopez-Lira (2023) machinery has no incremental predictive power. While these results do not rule out the possibility that managers issue more equity when analysts are overly optimistic, they also do not provide any support for this interpretation. These results simply show that low-EPS firms are more likely to issue equity.
4. **Overpricing and anomaly returns.** Binsbergen, Han, and Lopez-Lira (2023) also examine the association between predicted biases and anomaly returns. This result builds on Stambaugh, Yu, and Yuan's (2012) finding that anomalies' short legs are particularly profitable following periods of high sentiment. In short, the argument is that overvaluation is more difficult, costlier, and riskier to arbitrage than undervaluation and therefore more prevalent. Binsbergen, Han, and Lopez-Lira (2023) suggest that their estimated bias measure helps locate overpriced firms and, thereby, can enhance the profitability of various anomalies. Their estimates here are also not sensitive to the look-ahead bias; removing the look-ahead bias leaves the conditional alpha unchanged (1.63% with a t-value = 4.28 versus 1.84% with a t-value

[4] This finding is consistent with the market timing hypothesis of Hirshleifer and Jiang (2010).

of 4.66). In this case, the random forest machinery adds some value. A linear model produces an alpha of 1.68% (t-value = 3.93).

We believe that it is important to rectify the potential misconceptions created by Binsbergen, Han, and Lopez-Lira (2023). First, these results could significantly distort the allocation of resources within the profession. If a simple earnings-forecasting methodology sets a standard of a $t(\hat{\alpha})$ of 5.84 for trading strategies, there is less incentive to pursue ideas that fall short of this benchmark. Second, referees tasked with assessing the marginal contributions of new studies might dismiss papers on analysts' forecasts—that find weaker but correct results—if they benchmark them against the look-ahead-biased results. Third, researchers using the forecast outputs from Binsbergen, Han, and Lopez-Lira (2023) would perpetuate the original error. Put differently, researchers might overlook important findings, referees might hinder the publication of the findings that reach the journals, and studies building on the incorrect conclusions could propagate the original error.

1. Replication of Binsbergen, Han, and Lopez-Lira (2023)

1.1. Data and sample

We use data from the same four sources as Binsbergen, Han, and Lopez-Lira (2023):

1. Monthly stock-level data from CRSP.
2. Analysts' consensus earnings forecasts and actual earnings from IBES Unadjusted Summary and Unadjusted Detail Files.[5]
3. Fundamental data from Financial Ratios Suite by Wharton Research Data Services.
4. Real-time macroeconomic indicators from the Federal Reserve Bank of Philadelphia.

Binsbergen, Han, and Lopez-Lira (2023) also graciously provided the conditional bias estimates from the original study. Our main sample covers U.S. common stocks (CRSP share codes 10 and 11) listed on the NYSE, AMEX, and Nasdaq (exchange codes 1, 2, and 3), running from January 1984 through December 2019. Following BHL, we examine analysts' forecasts of annual earnings at the 1- and 2-year horizons (FPI = 1 and 2), and quarterly earnings forecasts at the 1-, 2-, and 3-quarter horizons (FPI = 6, 7, and 8). We detail the sample construction procedure in Internet Appendix B.

[5] We take the consensus forecasts from the IBES Summary file. IBES provides actual earnings data in two separate files: IBES Summary and IBES Detail. Approximately 2% of all observations (both quarterly and annual earnings) are found only in one of these files. We find that we best match the BHL sample by combining the two sources of actuals. We analyze this issue in detail in Internet Appendix G.

We follow BHL and use three sets of variables to predict firms' future earnings: firm characteristics, macroeconomic variables, and analysts' forecasts. The variable with the look-ahead bias is the first item on the list. We list all 75 variables in Internet Appendix A.

1. **Firm characteristics:**
   (a) Realized earnings in the last period ← **look-ahead biased**.
   (b) Monthly stock prices and returns from CRSP.
   (c) Sixty-seven financial ratios from the Financial Ratios Suite by Wharton Research Data Services. We fill missing values with industry medians using the Fama-French 49 industry classification.[6]
2. **Macroeconomic variables:**[7]
   (a) Consumption growth, defined as the log difference of consumption in goods and services.
   (b) GDP growth, defined as the log difference of real GDP.
   (c) Growth of industrial production, defined as the log difference of Industrial Production Index.
   (d) Unemployment rate.
3. **Analyst forecasts:**
   (a) Analysts' consensus quarterly and annual earnings-per-share (EPS) forecasts.

Because BHL model the term structure of expectations, they estimate models with different target variables: 1-quarter-ahead (Q1), 2-quarters-ahead (Q2), 3-quarters-ahead (Q3), 1-year-ahead (A1), and 2-years-ahead (A2) earnings. BHL define realized earnings and analyst forecasts on the list above to match the target variable. For example, when they forecast one-quarter-ahead earnings, "realized earnings" is the prior quarter's earnings and analysts' forecast is about 1-quarter-ahead earnings. On the other hand, when they forecast 2-years-ahead earnings, "realized earnings" is the 1-year-ahead earnings and the forecast is about 2-years-ahead earnings. The issue is that the 1-year-ahead earnings are unknown at the time of the forecast. Three of the five models—those with Q2, Q3, and A2 earnings as the target—have the look-ahead bias.

[6] Binsbergen, Han, and Lopez-Lira (2023) note in Internet Appendix A.2 that they "consider another twenty-six fundamental values per share derived from these financial ratios." They do not enumerate these additional per-share characteristics.

[7] The Federal Reserve Bank of Philadelphia issues monthly releases for consumption, real GDP, and industrial production, and quarterly releases for unemployment. We employ the latest vintages available each month.

Table 2
Replication details

Panel A: Choices experimented

	BHL	Replication
Timing of last-period realized earnings	Unspecified	**With look-ahead: $t+h-1$** Without look-ahead: $t-1$
Standardization	Unspecified	Z-scores Ranks **None**
Winsorization	Features (text) Features and targets (code)	Features **Features and targets**
Per-share data	Unspecified	Include **Exclude**
Sample fraction	1% (text) 5% (code)	1% **5%**
Feature fraction	Top five features (text) plus 50% at random (code)	Top five features plus 50% at random **All features**

Panel B: Other details

Feature timing	Analysts' forecast at time t and all other features at $t-1$
Number of trees	2,000
Maximum depth	7
Minimum node size	5
Training window	24 months for year-2 forecasts and 12 months for other forecasts

This table reports the specifications of the replication. Panel A reports which details are unspecified in BHL and the set of choices tried. The bolded option corresponds to the main specification, chosen on the basis of providing a better match with the results in BHL and/or simplicity. Panel B reports the other implementation details provided in BHL.

1.2. Model specification and implementation details

We train a random forest model each month t to predict firms' EPS for each horizon h:

$$E_t[AE_{i,t}^h]=f_t^h\left(\boldsymbol{X}_{i,t}|\boldsymbol{\gamma}\right), \tag{1}$$

where $AE_{i,t}^h$ denotes stock i's actual EPS at horizon h relative to time t, and f_t^h denotes the random forest model that takes the feature set $\boldsymbol{X}_{i,t}$ as an input with hyperparameters $\boldsymbol{\gamma}$. The forecasting horizon h takes the values of 1 quarter (Q1), 2 quarters (Q2), 3 quarters (Q3), 1 year (A1), and 2 years (A2).

We follow BHL's implementation. When the necessary details are missing, we search over plausible choices and select the ones that yield results that best match those in BHL.[8] When the results are very similar between different choices, we choose the simpler one.[9] We describe our specification details below and summarize them in Table 2.

[8] Although BHL provide replication code on the journal website, this code does not include the data processing steps that induce the look-ahead bias. Moreover, as we describe below, in some cases the code conflicts with what is stated in the study. The crucial step in our study is the tedious process of reverse-engineering the specification underneath the BHL results, thereby pinpointing the look-ahead bias.

[9] We present the results for alternative specifications in Internet Appendix F.1.

1. **Standardization**. Binsbergen, Han, and Lopez-Lira (2023) state in footnote 15 that they standardize the features. However, they do not describe the standardization procedure. We try cross-sectional Z-scores and ranks, as well as not standardizing. The replication results are similar under all these choices, with the unstandardized specification yielding slightly better-matching results.[10]
2. **Winsorization**. Binsbergen, Han, and Lopez-Lira (2023) state in footnote 15 that they winsorize the features at the 1% level. The code available on the journal's website also winsorizes the target variables. We experiment with only winsorizing the features and winsorizing both the features and targets. The latter yields more closely matching results.
3. **Per-share data**. Binsbergen, Han, and Lopez-Lira (2023) state in Internet Appendix A2 that they also consider 26 additional fundamental per-share characteristics, such as the book equity per share and the current debt per share to improve the forecasts. They do not, however, provide a full list of these per-share variables. We try both including and not including a set of per-share characteristics we deem plausible (see Internet Appendix Table A.1). The results are similar with and without these characteristics and we therefore exclude these variables from the main specification.
4. **Hyperparameters of the random forest regression**. A random forest regression is an ensemble of decision trees. Various hyperparameters, such as maximum depth and sample fraction, control what type of variation is induced in the data to generate each tree. The following choices are about these hyperparameters:

   (a) **Sample fraction**. Binsbergen, Han, and Lopez-Lira (2023) state in Table 1 that they use a sample fraction of 1% for training each decision tree. The code on the journal website uses a sample fraction of 5%. Using the 5% fraction produces a better match with the reported results.

   (b) **Feature fraction**. This parameter determines the number of features considered in each split. Binsbergen, Han, and Lopez-Lira (2023) state in Internet Appendix A1 that they always include the five most important features for every split from which to choose. In the code posted on the journal website, they appear to include also 50% of all available features at random. However, this line is commented out, in which case the code would default to drawing

[10] Unlike regressions with regularization penalties, such as LASSO (L1 norm) and ridge (L2 norm) regressions, random forest regressions do not present an inherent need for feature normalization. However, cross-sectional normalization can make a difference in panel data in random forest regressions by eliminating shifts in the features' distributions.

the square root of the total number of features.[11] We consider two implementations. First, we use R and estimate the random forest regression twice. In the first stage, we include all features; in the second stage, we include the top-five features from the first stage and 50% of all features at random. Second, we use Python and set the feature fraction to 1.0 so that every split considers all features and, therefore, the top-five features are always included as well. The two implementations yield very similar results and we therefore use the specification with all features.

(c) **Number of trees**. The size of the random forest is 2,000 trees.

(d) **Maximum depth**. Each tree has a maximum depth of seven.

(e) **Minimum node size**. The minimum number of samples required for further splitting a node is five.

5. **The timing of the features.** Binsbergen, Han, and Lopez-Lira (2023) state in Internet Appendix A3 that they require the forecasting variables to be available prior to the time when analysts' forecasts become available. Hence, when constructing the random forest forecasts in month t, we use analysts' forecasts available in month t and all other variables available at the end of month $t-1$.
6. **Training window**. We train the model monthly using a 12-month rolling window when forecasting quarterly and 1-year-ahead earnings. We use 24-month rolling windows to forecast 2-years-ahead earnings. These choices are the same as those reported in BHL.

1.3. Man versus machine earnings forecasts

Table 3 shows that we replicate BHL's sample and earnings forecasts closely. In panel A we report values from Table 2 in BHL. Panels B and C present our replication results with and without the look-ahead bias. Each panel has five rows, with each row corresponding to a different forecasting horizon. All earnings-related variables are in the units of earnings per share.

Columns 1, 2, and 3 report the sample averages of random forest forecasts (RF), analysts' forecasts (AF), and actual EPS (AE). Columns 4 and 5 report the average differences between the RF forecasts and actual earnings and the t-values associated with these differences. The AF-minus-AE differences in columns 6 and 7 indicate that analysts' forecasts are systematically upward

[11] The popular machine learning packages in neither Python nor R gives an out-of-the-box method for estimating a random forest regression in which every split includes the five most important features. To implement this scheme, the following steps need to be followed: (1) train an initial random forest model with feature fraction set to 1.0, (2) identify the top-five features based on feature importance, and (3) reestimate the random forest model with the option of always including the top-five features. The RANGER package in R, which the authors use in their code to estimate the random forest model, gives an ALWAYS.SPLIT.VARIABLES option of including a fixed set of features into every split.

Table 3
The term structure of earnings forecasts via machine learning

	(1) RF	(2) AF	(3) AE	(4) RF–AE	(5) t(RF–AE)	(6) AF–AE	(7) t(AF–AE)	(8) (RF–AE)^2	(9) (AF–AE)^2	(10) N
					Panel A: BHL original (Table 2)					
Q1	0.290	0.319	0.291	−0.000	−0.17	0.028	6.59	0.076	0.081	1,022,661
Q2	0.323	0.376	0.323	−0.001	−0.13	0.053	10.31	0.094	0.102	1,110,689
Q3	0.343	0.413	0.341	0.002	0.31	0.072	11.55	0.121	0.132	1,018,958
A1	1.194	1.320	1.167	0.027	1.64	0.154	6.24	0.670	0.686	1,260,060
A2	1.387	1.771	1.387	−0.004	−0.07	0.384	8.33	1.897	2.009	1,097,098
					Panel B: Replication with look-ahead bias					
Q1	0.288	0.317	0.289	−0.001	−0.37	0.028	6.50	0.071	0.078	1,024,654
Q2	0.319	0.374	0.323	−0.004	−0.83	0.051	10.17	0.087	0.096	1,119,537
Q3	0.338	0.411	0.341	−0.002	−0.40	0.070	11.38	0.111	0.125	1,026,893
A1	1.184	1.306	1.157	0.027	1.89	0.149	6.14	0.602	0.655	1,265,472
A2	1.349	1.727	1.359	−0.011	−0.20	0.368	8.26	1.643	1.856	1,136,233

Panel C: Replication without look-ahead bias and accuracy comparison

					Accuracy improvement in (RF–AE)^2			
	Replication without look-ahead bias				Against look-ahead-biased forecast		Against linear model forecast (Hughes, Liu, and Su 2008)	
	RF	RF–AE	t(RF–AE)	(RF–AE)^2	Difference	t-value	Difference	t-value
Q1	0.288	−0.001	−0.37	0.071			0.002	1.30
Q2	**0.318**	**-0.004**	**-0.95**	**0.090**	−0.003	−6.48	0.003	1.10
Q3	**0.337**	**-0.003**	**-0.51**	**0.120**	−0.009	−6.84	0.004	0.91
A1	1.184	0.027	1.89	0.602			−0.005	−0.63
A2	**1.356**	**-0.004**	**-0.07**	**1.805**	−0.161	−10.50	−0.003	−0.05

Panels A and B report time-series averages of machine learning EPS forecasts (RF), analysts' EPS forecasts (AF), and realized EPS (AE), and the time-series averages of the cross-sectional mean-squared forecasting errors by random forest, $(RF-AE)^2$, and by analysts, $(AF-AE)^2$. N denotes the number of observations. We also report the Newey and West (1987) adjusted t-values of the differences between earnings forecasts and realized earnings. The standard errors for quarterly forecasts are adjusted with 3 lags and those for annual forecasts are adjusted with 12 lags. Panel A reports estimates from Table 2 in BHL. Panel B shows the replication results with the look-ahead bias. The first four columns in panel C report the replication results without the look-ahead bias. The look-ahead bias affects the bolded estimates. The next two columns estimate the accuracy improvement from the look-ahead bias. Accuracy improvement is the time-series difference in the cross-sectional mean-squared forecasting errors. The last two columns estimate the accuracy improvement over an OLS model that uses Hughes, Liu, and Su's (2008) variables and analysts' forecasts for the corresponding horizon. The t-values use Newey-West adjusted standard errors. The sample starts in January 1986 and ends in December 2019.

biased; the differences are positive and statistically significant at all horizons. The random forest-based forecasts in column 4, by contrast, are close to zero.

We successfully replicate these estimates. The estimates in panel B (with a look-ahead bias) are close to the BHL numbers in panel A.[12] The mean-squared forecasting errors in columns 8 and 9 measure how accurate the random forest model is relative to analysts. BHL's result is that, by unbiasing analysts' forecasts, the random forest model's accuracy consistently exceeds that of the analysts. The results in panel B confirm this finding.

Panel C shows that removing the look-ahead bias substantially reduces the model's accuracy at the affected horizons (Q2, Q3, and A2). Both in BHL and in our replication, the model with a look-ahead bias has the greatest advantage over analysts in the task of predicting 2-years-ahead earnings. Panel B shows that, at this horizon, the look-ahead-biased model decreases mean-squared errors by 0.213 ($=1.856-1.643$) units—11.5%—from the analyst benchmark. However, panel C shows that removing the look-ahead bias shrinks this gap to 0.051 ($=1.856-1.805$). That is, we attribute three-quarters of the random forest model's superior performance to the look-ahead bias. This degradation estimate is significant, with a t-value of -10.50. The degradations at the two other affected horizons (2- and 3-quarters-ahead forecasts) are 33.3% and 64.3%. Nevertheless, consistent with BHL, the machine learning–based forecasts are more accurate than analysts at all five horizons even absent the look-ahead bias.

How significantly does the random forest technique enhance forecast accuracy? In panel C's rightmost columns, we compare the look-ahead-bias-free random forest estimates to a linear model. We construct OLS forecasts using analysts' forecasts and the eight variables from Hughes, Liu, and Su (2008).[13] The accuracy improvement of the random forest forecasts over the OLS forecasts is economically small and statistically insignificant at all horizons. What matters more than the technique is including analyst forecasts as one of the features.[14] Removing this feature from BHL or the linear model significantly degrades their performance. Most prior studies, such as So (2013), have predicted earnings without including analyst forecasts as an input to test whether analysts outperform simple statistical models.[15]

[12] Panel B's replication estimates imply slightly higher accuracy. Despite our best efforts, we could not identify the source of this discrepancy. In particular, by using 2020 vintages of the IBES and CRSP data, we confirmed that this discrepancy is not due to the changes made to the IBES and CRSP data since that date.

[13] The eight variables are accruals, analysts' long-term EPS growth forecast, long-term sales growth rate in the past 5 years, change of PP&E, change of long-term assets, past earnings surprise, past-12-month stock return, and analysts' forecast revision.

[14] Binsbergen, Han, and Lopez-Lira (2023) note on p. 2367 that a model that includes analyst forecasts as a feature should, in practice, be at least as accurate as the analysts.

[15] These findings are consistent with Campbell et al. (2024), who find that the accuracy of machine learning forecasts is highly sensitive to the machine learning model specification. In their analysis, 90% of the models fail to beat analysts.

Table 4
Portfolios sorted on conditional bias

	Conditional bias (average BE) quintile					
	1	2	3	4	5	5–1
BHL original (Table 5)	1.32	0.98	0.79	0.47	−0.14	−1.46
	(6.53)	(4.53)	(3.18)	(1.62)	(−0.35)	(−5.11)
Replication with look-ahead bias	1.38	0.92	0.61	0.24	−0.21	−1.59
	(6.74)	(4.23)	(2.44)	(0.78)	(−0.51)	(−5.08)
Replication without look-ahead bias	0.99	0.90	0.91	0.85	0.75	−0.25
	(4.83)	(4.10)	(3.62)	(2.78)	(1.81)	(−0.80)

This table reports time-series average returns (in percent) for value-weighted monthly rebalanced portfolios formed on analysts' average conditional earnings forecast bias quintiles. The first and second rows are from Table 5 in BHL. The remaining four rows report our replication results with and without the look-ahead bias. The t-values (in parentheses) use White's (1980) heteroskedasticity robust standard errors, as in BHL. The sample starts in January 1986 and ends in December 2019.

A reader might suggest that the comparison between the Hughes, Liu, and Su (2008) and BHL models is unfair because of a hindsight bias: whereas the BHL model has to learn from the data which variables are useful for predicting earnings, Hughes, Liu, and Su (2008) may have tried different sets of predictors and chosen the final set to maximize the model's in-sample accuracy. We can remove any such bias by comparing the two models in the data that have accrued subsequent to the Hughes, Liu, and Su (2008) sample. We find that the differences between the two models are statistically insignificant at all forecast horizons in both the pre- and post-2006 data.[16] The pool of 10 test statistics range from −0.82 to 1.27, with the largest t-value being for the Q3 horizon in the post-2006 data. The variable-selection bias, even if one were to exist in the original study, therefore cannot explain why the linear model performs as well as the random forest model.[17]

1.4. Conditional bias and stock returns

Table 4 presents our replication of BHL's tests that predict the cross section of stock returns. At each month t, we measure the biases in analysts' forecasts as the differences between the analyst (AF) and random forest (RF) forecasts, scaled by the month $t-1$ closing stock price:

$$Bias_{i,t}^{h} = \frac{AF_{i,t}^{h} - RF_{i,t}^{h}}{Price_{i,t-1}}, \tag{2}$$

where $AF_{i,t}^{h}$ and $RF_{i,t}^{h}$ are analysts' and random forest forecasts in month t for firm i's h-period-ahead EPS. We average the biases across all horizons and use this "Average BE" to predict returns.[18] The trading strategy sorts stocks into

[16] The sample period in Hughes, Liu, and Su (2008) ends in 2006.

[17] We discuss hindsight bias in more detail in Internet Appendix E.3.

[18] We follow BHL and compute "Average BE" only for firm-month observations in which the bias estimate is nonmissing for at least two of the five horizons.

quintiles by the average BE each month and buys (sells) stocks with high (low) estimated bias. This process is identical to that in BHL.

The first two rows of Table 4 show the original results from BHL. The low-bias (quintile 1) portfolio earns an average return of 1.32% per month with a t-value of 6.53, the high-bias portfolio earns an average of -0.14% per month (t-value = -0.35), and the difference between the two has a t-value of -5.11. Our replication with a look-ahead bias produces an equally profitable strategy. The difference between the top and bottom quintiles is -1.59% (t-value = -5.08).

The BHL strategy is exceptionally profitable given that the sample runs from 1986 through 2019 (many anomalies are weaker in post-2000 data) and leans towards larger stocks by the virtue of requiring analyst coverage for multiple horizons (many anomalies are weaker among larger firms).[19] Figure 1 shows the distribution of the annualized Sharpe ratios of 151 anomalies in Chen and Zimmermann's (2022) database. We make each anomaly comparable with the BHL strategy by using consistent samples and portfolio rules. The Sharpe ratio of the BHL strategy is 0.95.[20] Compared to all other anomalies in Chen and Zimmermann's (2022) database, this Sharpe ratio is the highest by a wide margin; the second highest estimate in this pool of anomalies is 0.78.

Most of the profits of this trading strategy, however, stem from the look-ahead bias. Table 4 shows that, when we remove the look-ahead bias, the difference between the top and bottom quintiles decreases by four-fifths to -25 basis points. This average return is economically small and statistically indistinguishable from zero. The red vertical line in Figure 1 shows that, without the look-ahead bias, the BHL strategy earns a Sharpe ratio of 0.15.

Figure 2 shows cumulative returns for three long-short strategies that trade against estimated biases. The blue line uses the estimates from BHL and the orange dashed line the estimates from the replication with a look-ahead bias. The two strategies move in lockstep. The correlation between the two over the full sample period is 0.95. The replication without the look-ahead bias (green dotted line), by contrast, is consistently unprofitable.

In Table 5, we control for common risk factors. Panels A, B, and C show BHL's original results and the replication results with and without the look-ahead bias. Panels A and B show that both the BHL strategy and our replication with the look-ahead bias earn large and statistically highly significant alphas under the CAPM and the Fama-French three- and five-factor models. Our replication closely matches BHL's estimates in terms of both alphas and factor loadings. Panel C shows that, after removing the look-ahead bias, the profits attenuate. While the strategy's CAPM and three-factor model alphas remain statistically significant, this remaining profitability is due to the correlations

[19] See, for example, Green, Hand, and Zhang (2017) and Hou, Xue, and Zhang (2020).

[20] We use the conditional bias data provided by BHL to compute their strategy's Sharpe ratio.

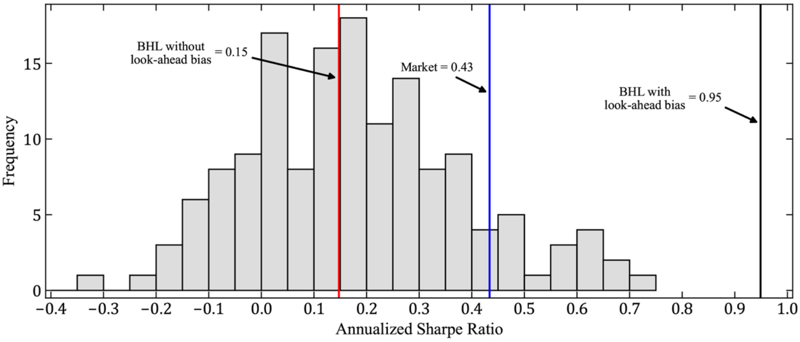

Figure 1
Sharpe ratios of 151 anomaly portfolios
This figure plots the distribution of annualized Sharpe ratios of 151 anomalies from Chen and Zimmermann's (2022) database. We impose the following restrictions both on the BHL strategy and all anomalies. The sample consists of common shares traded in NYSE, AMEX, and Nasdaq that have a positive book value of equity in the prior year and a share price above one dollar in the prior month. We also exclude microcap stocks, defined as those with a market value of equity below the 20th NYSE percentile in the prior month. We require the anomaly variables to be continuous (that is, excluding discrete and categorical variables) and available from 1986 to 2019 for at least 500 stocks in the average month. We augment the Chen-Zimmermann database with the price, size, and short-term reversal signals that it omits. All anomalies are value-weighted long-short portfolios based on monthly rebalanced quintiles. The black, blue, and red vertical lines indicate the Sharpe ratios of the BHL strategy (with the look-ahead bias), the market portfolio, and our replication of BHL (without the look-ahead bias), respectively. The Sharpe ratio of the market portfolio is computed from July 1926 to December 2019.

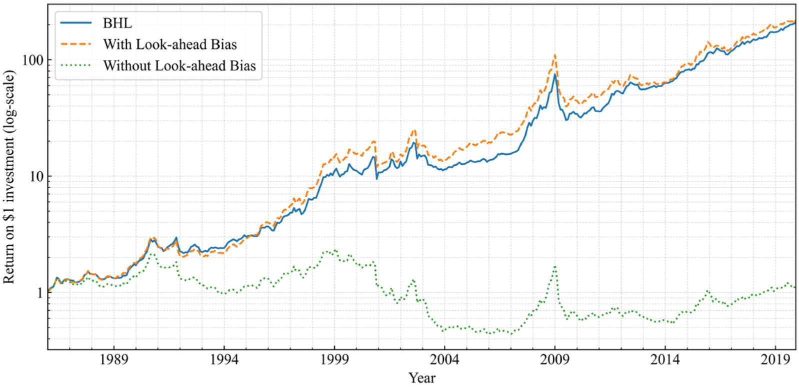

Figure 2
Cumulative returns of value-weighted portfolios formed on conditional biases
This figure shows cumulative returns for monthly rebalanced value-weighted strategies that buy (sell) stocks in the lowest (highest) "Average BE" quintile. We implement this strategy using (a) the estimates provided by BHL (blue solid line), (b) our replication with a look-ahead bias (orange dashed line), and (c) our replication without a look-ahead bias (green dotted line). The sample period begins in January 1986 and ends in December 2019.

with the investment and profitability factors—analysts are more positive about unprofitable stocks that invest aggressively (Cooper, Gulen, and Schill 2008; Bouchaud et al. 2019). The five-factor model alpha of −0.41% per month

Table 5
Time-series tests with common asset pricing models

	CAPM		FF3		FF5	
	Coef.	t-value	Coef.	t-value	Coef.	t-value
Panel A: BHL original (Table 6)						
Intercept	−1.85	−7.18	−1.96	−8.64	−1.54	−5.84
MKT	0.56	7.53	0.53	7.86	0.38	5.28
SMB			0.80	7.06	0.61	5.17
HML			0.58	5.25	0.95	7.12
RMW					−0.68	−4.10
CMA					−0.53	−1.93
Panel B: Replication with look-ahead bias						
Intercept	−2.00	−7.25	−2.10	−8.63	−1.71	−6.07
MKT	0.61	7.49	0.56	7.56	0.43	5.71
SMB			0.88	7.22	0.66	5.14
HML			0.62	5.35	0.90	6.27
RMW					−0.75	−4.01
CMA					−0.30	−1.09
Panel C: Replication without look-ahead bias						
Intercept	−0.67	−2.45	−0.76	−3.26	−0.41	−1.54
MKT	0.61	7.75	0.55	7.83	0.44	5.94
SMB			0.96	7.82	0.75	5.69
HML			0.61	5.48	0.84	5.90
RMW					−0.72	−3.83
CMA					−0.18	−0.67

This table reports the alphas and factor loadings for the value-weighted long-short portfolio under the CAPM, Fama and French (1993) three-factor (FF3), and Fama and French (2015) five-factor models (FF5). The results in panel A are from BHL's Table 6. Panels B and C present the results using our replicated average conditional bias (BE) with and without look-ahead bias. The t-values use White's (1980) heteroskedasticity robust standard errors, as in BHL. The sample starts in January 1986 and ends in December 2019.

(t-value = −1.54) shows that, net of these factor exposures, a strategy that trades against the estimated biases is not abnormally profitable. The profits generated by the BHL strategy are therefore not unique to the information derived from analysts' forecasts; the factors of the five-factor model do not, after all, use such information and yet these factors span the BHL strategy.

In Table 6, we measure the performance of the strategy (without look-ahead bias) under four alternative factor models: the Fama-French-Carhart six-factor model (FFC6, Fama and French 2015; Carhart 1997), Hou, Xue, and Zhang's (2015) q-factor model, Hou et al.'s (2021) q5-factor model (HMXZ), Stambaugh and Yuan's (2016) mispricing factor model (SY), and Daniel, Hirshleifer, and Sun's (2020) short- and long-horizon behavioral model (DHS). The strategy earns no abnormal profits vis-á-vis any of these models.[21]

Binsbergen, Han, and Lopez-Lira (2023) report average returns and alphas for horizon-specific trading strategies in their Internet Appendix. Each of these strategies bets against the estimated bias at one of the five horizons. They find highly significant alphas only for the three horizons affected by the look-ahead bias: Q2, Q3, and A2. In Table C.3 in the Internet Appendix, we report

[21] In Internet Appendix Tables C.1 and C.2, we show that the same conclusion holds under alternative specifications such as equal-weighted portfolios and Fama-MacBeth regressions.

Table 6
Alternative models without look-ahead bias

Model	FFC6	HXZ-$q4$	HMXZ-$q5$	SY	DHS
Intercept	−0.02	−0.08	−0.01	0.07	0.31
	(−0.11)	(−0.27)	(−0.02)	(0.25)	(0.87)

This table reports the alphas of the long-short strategy without look-ahead bias under the Fama and French (2015) and Carhart (1997) six-factor model (FFC6), Hou, Xue, and Zhang (2015) q-factor model (HXZ), Hou et al. (2021) $q5$-factor model (HMXZ), Stambaugh and Yuan (2016) mispricing factor model (SY), and Daniel, Hirshleifer, and Sun (2020) short- and long-horizon behavioral model (DHS). The t-values (in parentheses) use White's (1980) heteroskedasticity robust standard errors as in BHL. The sample starts in January 1986 and ends in December 2019.

average returns and alphas for the same horizon-specific strategies. These by-horizon results are similar to the average-bias results: some of the strategies earn significant CAPM and three-factor model alphas, but none of them earn statistically significant average returns or alphas under modern multifactor models.

Is the trading strategy without the look-ahead bias economically meaningful? Another benchmark for this question is to measure the strategy's profitability relative to a nearly identical strategy explored in So (2013); similar to BHL, So (2013) forecasts earnings, measures biases in analysts' forecasts, and trades against these estimated biases. The difference between So (2013) and BHL is in the choice of the forecasting technique (OLS versus a random forest model) and the set of features. Binsbergen, Han, and Lopez-Lira (2023) espouses this comparison, presenting a detailed breakdown of how their strategy performs relative to So (2013).

We reproduce and extend this comparison in Table 7. The estimates on the first and fourth rows are from BHL. Binsbergen, Han, and Lopez-Lira (2023) note that while the So (2013) strategy is highly profitable in the full sample, (1) it is significantly less profitable than the random forest–based strategy and (2) it attains, at best, borderline statistical significance in the post-2000 sample. The pre-post comparison indicates that while So's (2013) strategy decays significantly, theirs does not. Our replications on the other rows of Table 7 show that BHL represents an improvement over So (2013) only because of the look-ahead bias. Absent this bias, the strategy is less profitable than the So (2013) strategy both in the full sample and each of the subsamples.

1.5. Pinpointing the look-ahead bias

The estimates above suggest that the key results in BHL stem from a look-ahead bias. Our inference, however, is somewhat indirect: (1) we successfully replicate BHL's results when we build in the look-ahead bias and (2) removing this bias markedly changes them. Because the authors' code on the journal website does not include the data-processing steps—which create the look-ahead bias—it is instructive to consider a test that can more directly identify the look-ahead bias.

Table 7
Portfolios sorted on conditional bias: So (2013) versus Binsbergen, Han, and Lopez-Lira (2023)

		Full	Subsample			
Strategy	Specification	Sample	Pre	Post	Pre−Post	p-value
BHL	Original	−1.46 (−5.11)	−1.67 (−4.52)	−1.30 (−3.09)	−0.37	0.51
	Replication with look-ahead bias	−1.59 (−5.08)	−1.82 (−4.63)	−1.42 (−3.14)	−0.40	0.53
	Replication w/o look-ahead bias	−0.25 (−0.80)	−0.52 (−1.35)	−0.06 (−0.13)	−0.46	0.47
So (2013)	BHL's replication	−0.67 (−4.05)	−1.11 (−3.79)	−0.33 (−1.79)	−0.78	0.03

This table reports average returns for value-weighted monthly rebalanced long-short portfolios formed on various measures of the earnings forecast bias. The estimates on rows 1 and 4 are from (Binsbergen, Han, and Lopez-Lira 2023, Table A13); those on rows 2 and 3 are from our replication of BHL with and without the look-ahead bias. We report the estimates for the full, pre-2000, and post-2000 samples. The t-values use White's (1980) heteroskedasticity robust standard errors as in BHL. The two rightmost columns test the difference in average returns between the pre- and post-2000 subperiods.

Given the nature of the look-ahead bias, we can pin it down by regressing future earnings against BHL's forecasts:

$$AE^h_{i,t} = \alpha + \beta_1 FE^h_{i,t} + \beta_2 FE^{h+1}_{i,t} + \underbrace{\gamma_1 AF^h_{i,t} + \gamma_2 AF^{h+1}_{i,t}}_{\text{Controls}} + \varepsilon_{i,t}, \tag{3}$$

where $AE^h_{i,t}$ is the horizon-h realized earnings; $FE^h_{i,t}$ and $FE^{h+1}_{i,t}$ are the random forest forecasts for horizon-h and horizon-$h+1$ earnings (provided by BHL); and $AF^h_{i,t}$ and $AF^{h+1}_{i,t}$ are analysts' forecasts as control variables. The regression coefficients β_1 and β_2 represent the marginal effects of the random forest forecasts when controlling for analyst forecasts. The key parameter of interest is β_2. If FE^{h+1} is indeed a function of AE^h—that is, there is a look-ahead bias—then β_2 should be positive even when controlling for FE^h. Put differently, if there is the hypothesized look-ahead bias, Equation (3) can be rewritten as

$$AE^h_{i,t} = \alpha + \beta_1 \hat{f}^h_t(AE^{h-1}_{i,t}, \text{â¦}) + \beta_2 \hat{f}^{h+1}_t(AE^h_{i,t}, \text{â¦}) + \underbrace{\gamma_1 AF^h_{i,t} + \gamma_2 AF^{h+1}_{i,t}}_{\text{Controls}} + \varepsilon_{i,t}, \tag{4}$$

where $\hat{f}^h_t(\cdot)$ is the horizon-h random forest trained at time t and we highlight in red the target variable's appearance on the right-hand side as part of the horizon $h+1$ forecast.

We report the regression results in Internet Appendix D and summarize them here. We first estimate the regressions using the forecast data BHL provided. We find that the $\hat{\beta}_2$ estimates are significantly positive with t-values as high as 30.2, while the $\hat{\beta}_1$ estimates are negative. This $\hat{\beta}_2$ pattern is exactly what Equation (4) predicts. When we replicate the machine-learning models with the hypothesized look-ahead bias, we get similar estimates. However, when we remove the look-ahead bias, this pattern disappears.

2. The Effect of the Look-Ahead Bias on the Equity Issuance and Anomaly Return Results

In addition to the forecast accuracy and return predictability results, BHL find that:

(a) Managers of firms with high conditional bias issue more equities (Section 4.4, Table 9).

(b) Stocks with high conditional bias tend to be shorted by anomalies (Section 4.3, Table 7).

(c) Anomalies are stronger among stocks with higher conditional biases (Table 8).

Since the look-ahead bias contributes significantly to the accuracy and return predictability results—absent this bias, linear models display performance comparable to random forests—it is important to re-examine these other results as well. We evaluate these results from three perspectives:

1. *Sensitivity to the look-ahead bias*: Are the results sensitive to the look-ahead bias?
2. *Interpretation*: Are the original economic interpretations supported by the data?
3. *Marginal contribution of machine learning techniques*: Do machine learning models outperform linear models?

We summarize our findings here and present the detailed results in Internet Appendix E. First, we find that the results (a), (b), and (c) are not very sensitive to the look-ahead bias; that is, although the estimates change, they retain their signs and statistical significance when we remove the bias. However, both the economic interpretations of the results and inferences about the marginal contribution of the machine learning technique change.

The interpretations of results (a) and (b) do not appear to be justified by the data. To see why, it is useful to decompose the authors' key variable into its two constituents, Conditional Bias = Analyst Forecast − Expected Earnings Benchmark. A high conditional bias may reflect high analyst forecasts or firms having low expected earnings. When the authors sort stocks into portfolios by this measure (or include it as a regressor), it could be that it is indeed the difference that matters—or it could be that conditional bias "works" because it is a noisy transformation of expected earnings (the second component). This distinction is important. If we obtain the same estimates without using any analyst information, then the results cannot be evidence of how biases in these forecasts correlate with management behavior.

We reproduce, replicate, and extend BHL's equity issuance results in Internet Appendix Table E.1. We successfully replicate the key estimates and find that these estimates remain largely unchanged when we remove the look-ahead bias. However, replacing the conditional bias measure with expected earnings increases the economic and statistical significance of the estimates, and replacing the random forest predictions with the realized earnings—which is the random-walk forecast often employed in the accounting literature[22] —renders these results stronger still. In multivariate regressions, the expected earnings benchmark subsumes the conditional bias measure's predictive power. Put differently, we obtain the same results as BHL without using any information on analysts or even using any techniques to forecast earnings. This irrelevancy result casts doubt on the interpretation that conditional bias relates to managers' stock issuance behaviors. These results simply show that low-EPS firms are more likely to issue equity. Although managers may pay heed to how overly optimistic analysts are, the results in BHL do not establish such a correlation.

Result (b)—that stocks shorted by many asset pricing anomalies are predominantly stocks with high conditional biases—has the same interpretational ambiguity. We replicate this result with the look-ahead bias but also find that the estimates are almost identical when we remove this bias (Internet Appendix Table E.2). However, similar to the stock issuance result, it is the forecasted earnings component that drives the result. Sorting stocks into portfolios by this benchmark alone creates a similar spread in stocks' locations across the portfolios.

The remaining result in BHL—that anomalies are stronger among stocks with high conditional biases—is neither affected by the look-ahead bias nor subject to ambiguity about its interpretation. We successfully replicate BHL's result when we build in the look-ahead bias, and find that removing this bias leaves the estimates largely unchanged. In this case, we find that the random forest technology adds some value: conditioning on the bias constructed with the random forest model (t-value = 4.66) outperforms the one built using the linear model (t-value = 3.96) (Internet Appendix Table E.3).

3. Alternative Machine Learning Forecasts

The key results in BHL do not benefit from the use of the random forest technique. Linear models produce forecasts and trading strategies of equal accuracy and superior profitability. These findings, however, leave open the possibility that alternative techniques might perform better. In Section F of the Internet Appendix, we construct alternative random forest specifications and employ other machine learning techniques to forecast earnings. When we examine the performance of these techniques through the lens of the trading strategy result, we find that none of these alternative specifications brings back the alpha.

Although we find no evidence that machine learning–based estimates of conditional biases generate useful alpha signals, the premise that machine

[22] See, for example, Bradshaw et al. (2012).

learning may improve forecasting accuracy is appealing. We construct expected earnings benchmarks using OLS, partial least squares, LASSO, elastic net, random forest, and LightGBM, and also build a composite benchmark that averages these forecasts. In this analysis, we use the same features as BHL instead of, for example, the smaller set of So (2013) variables. The composite model generates the most accurate forecasts overall. This finding is consistent with the basic tenet of machine learning that ensemble techniques—or, in this case, an ensemble of ensemble techniques—often outperform individual models.[23] Although analysts' biases measured against these benchmarks neither significantly predict stock returns nor are more accurate than simpler linear models, they may still be useful in certain applications. As such, we have made our expected earnings data available on our website.[24]

4. Discussion and Conclusions

We revisit Binsbergen, Han, and Lopez-Lira's (2023) findings on the efficacy of machine learning models in unbiasing analyst forecasts and extracting estimates of conditional biases. A look-ahead bias drives the results on forecasting accuracy and alphas; the model that BHL train includes future earnings in the set of features. Removing this look-ahead bias, the model's accuracy declines substantially (it is no more accurate than a linear model) and the trading strategy earns no alpha under any modern factor model (it falls short of So's [2013] strategy built on the same idea).

Binsbergen, Han, and Lopez-Lira (2023) have significantly influenced the literature.[25] Their work is influential because the arguments are convincing, the empirical evidence strong, and the conclusions plausible. Their results and conclusions have been cited and techniques adopted by numerous other studies. We conclude by discussing where this area of research stands after adjusting the inferences for the look-ahead bias in BHL and propose future directions.

New alphas. Removing the look-ahead bias from BHL's trading strategy is important for two reasons. First, new results are not judged merely by their statistical significance but also by where they fall in the distribution of $t(\hat{\alpha})$s of the factor zoo (Harvey, Liu, and Zhu 2016). The trading strategy in BHL was an outlier, thereby affecting how other anomalies are perceived by readers. Second, our results show that even with machine learning techniques,

[23] Random forest, for example, is an ensemble technique: it is the average of simple regression trees generated by bootstrapping and randomizing the data and the aspects of feature selection.

[24] These data are available in a GitHub repository.

[25] As of March 2025, it has been cited 129 times according to Google Scholar. Several published papers apply Binsbergen, Han, and Lopez-Lira's (2023) model and/or alternative machine learning models to construct earnings expectations (e.g., Cassella et al. 2022; Silva and Thesmar 2024) and return expectations (Cao, Jiang et al. 2024). A number of working papers apply machine learning models to study related topics (e.g., Campbell et al. 2024; Cao, Guo et al. 2024).

predicting abnormal returns via analysts' biases is difficult. Machine learning techniques are not inherently superior to classical statistical approaches; although these techniques can undoubtedly add value, a blind switch from a classical technique to a machine learning technique does not guarantee automatic gains. We caution against the interpretation that (1) producing new alphas is easy with today's technology, and (2) ex-ante measures of analysts' biases strongly predict returns. Thus, the enterprise of discovering and understanding new alphas may still be of considerable value to the profession.

The term structure of earnings expectations and stock returns. The failure of BHL's measure to produce novel alpha does not necessarily imply that the term structure of analysts' expectation biases is unrelated to valuations; it may just be that we need better methods—machine learning or otherwise—to uncover this relationship. The recent literature has become keenly interested in how agents form beliefs over multiple horizons.[26] To understand the association with stock returns, there may be a benefit from modeling the entire term structure of expectations instead of trying to collapse this information into one number.

From correlation to causality. Even if a machine learning technique such as that in BHL predicts returns, the source of this predictability might be unclear; it could stem from either mispricing or risk. While classical statistical methods are used for inference, the primary goal of machine learning methods is high-quality predictions (Hastie et al. 2009). If we add a large number of features into a machine learning model to predict stock returns, it can be difficult to interpret the causal relationships and understand the economic rationale. More work is needed to understand the causal links (or the lack thereof) between analysts' forecast errors and stock returns.

We consider these areas as of particular interest to researchers focused on forecasting earnings and understanding the causal relation between analyst forecasts and stock returns. We do not intend this list to exclude the possibility that other avenues may prove as or even more fruitful.

Code Availability: The replication code is available in the Harvard Dataverse at: https://doi.org/10.7910/DVN/YWZJOP.

References

Binsbergen, J. H., X. Han, and A. Lopez-Lira. 2023. Man versus machine learning: The term structure of earnings expectations and conditional biases. *Review of Financial Studies* 36:2361–96.

Bouchaud, J.-P., P. Krüger, A. Landier, and D. Thesmar. 2019. Sticky expectations and the profitability anomaly. *Journal of Finance* 74:639–74.

Bradshaw, M. T., M. S. Drake, J. N. Myers, and L. A. Myers. 2012. A re-examination of analysts' superiority over time-series forecasts of annual earnings. *Review of Accounting Studies* 17:944–68.

[26] See, for example, Silva and Thesmar (2024) and Dessaint, Foucault, and Frésard (2024).

Campbell, J. L., H. Ham, Z. Lu, and K. Wood. 2024. Expectations matter: When (not) to use machine learning earnings forecasts. Working Paper, University of Georgia.

Cao, S., N. Guo, H. Xiao, and B. Yang. 2024. Can machines understand human skills? Insights from analyst selection. Working Paper, University of Maryland.

Cao, S., W. Jiang, J. Wang, and B. Yang. 2024. From man vs. machine to man + machine: The art and AI of stock analyses. *Journal of Financial Economics* 160:103910.

Carhart, M. M. 1997. On persistence in mutual fund performance. *Journal of Finance* 52:57–82.

Cassella, S., B. Golez, H. Gulen, and P. Kelly. 2022. Horizon bias and the term structure of equity returns. *Review of Financial Studies* 36:1253–88.

Chen, A. Y., and T. Zimmermann. 2022. Open source cross-sectional asset pricing. *Critical Finance Review* 11:207–64.

Cooper, M. J., H. Gulen, and M. J. Schill. 2008. Asset growth and the cross-section of stock returns. *Journal of Finance* 63:1609–51.

Daniel, K., D. Hirshleifer, and L. Sun. 2020. Short- and long-horizon behavioral factors. *Review of Financial Studies* 33:1673–736.

Dessaint, O., T. Foucault, and L. Frésard. 2024. Does alternative data improve financial forecasting? The horizon effect. *Journal of Finance* 79:2237–87.

Fama, E. F., and K. R. French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33:3–56.

———. 2015. A five-factor asset pricing model. *Journal of Financial Economics* 116:1–22.

Green, J., J. R. Hand, and X. F. Zhang. 2017. The characteristics that provide independent information about average us monthly stock returns. *Review of Financial Studies* 30:4389–436.

Harvey, C. R., Y. Liu, and H. Zhu. 2016. … and the cross-section of expected returns. *Review of Financial Studies* 29:5–68.

Hastie, T., R. Tibshirani, J. H. Friedman, and J. H. Friedman. 2009. *The elements of statistical learning: Data mining, inference, and prediction*. 2nd ed. New York: Springer.

Hirshleifer, D., and D. Jiang. 2010. A financing-based misvaluation factor and the cross-section of expected returns. *Review of Financial Studies* 23:3401–36.

Hou, K., H. Mo, C. Xue, and L. Zhang. 2021. An augmented q-factor model with expected growth. *Review of Finance* 25:1–41.

Hou, K., C. Xue, and L. Zhang. 2015. Digesting anomalies: An investment approach. *Review of Financial Studies* 28:650–705.

———. 2020. Replicating anomalies. *Review of Financial Studies* 33:2019–133.

Hughes, J., J. Liu, and W. Su. 2008. On the relation between predictable market returns and predictable analyst forecast errors. *Review of Accounting Studies* 13:266–91.

Kothari, S., E. So, and R. Verdi. 2016. Analysts' forecasts and asset pricing: A survey. *Annual Review of Financial Economics* 8:197–219.

La Porta, R. 1996. Expectations and the cross-section of stock returns. *Journal of Finance* 51:1715–42.

Newey, W. K., and K. D. West. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55:703–8.

Silva, T., and D. Thesmar. 2024. Noise in expectations: Evidence from analyst forecasts. *Review of Financial Studies* 37:1494–537.

So, E. C. 2013. A new approach to predicting analyst forecast errors: Do investors overweight analyst forecasts? *Journal of Financial Economics* 108:615–40.

Stambaugh, R. F., J. Yu, and Y. Yuan. 2012. The short of it: Investor sentiment and anomalies. *Journal of Financial Economics* 104:288–302.

Stambaugh, R. F., and Y. Yuan. 2016. Mispricing factors. *Review of Financial Studies* 30:1270–315.

White, H. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48:817–38.