# High-Throughput Asset Pricing

Andrew Y. Chen
Federal Reserve Board

Chukwuma Dim
George Washington University

June 2025*

**Abstract**

We apply empirical Bayes (EB) to mine data on 136,000 long-short strategies constructed from accounting ratios, past returns, and ticker symbols. This "high-throughput asset pricing" matches the out-of-sample performance of top journals while eliminating look-ahead bias. Naively mining for the largest Sharpe ratios leads to similar performance, consistent with our theoretical results, though EB uniquely provides unbiased predictions with transparent intuition. Predictability is concentrated in accounting strategies, small stocks, and pre-2004 periods, consistent with limited attention theories. Multiple testing methods popular in finance fail to identify most out-of-sample performers. High-throughput methods provide a rigorous, unbiased framework for understanding asset prices.

**JEL Classification**: G0, G1, C1

**Keywords**: stock market predictability, stock market anomalies, p-hacking, multiple testing

*Alternative title: "How I learned to stop worrying and love data mining." First posted to arXiv.org: November 2023. Replication code: `https://github.com/chenandrewy/high-throughput-ap`. Long-short strategy data: `https://sites.google.com/site/chenandrewy/`. We thank our discussants: Rudiger Weber, Julio Crego, and Morad Zekhnini for helpful comments. We also thank participants at the TBEAR Network Workshop (KIT), FutFin-Info, MFA, Iowa State, and Baylor. The views expressed herein are those of the authors and do not necessarily reflect the position of the Board of Governors of the Federal Reserve or the Federal Reserve System.

# 1 Introduction

Data mining refers to searching data for interesting patterns. This search leads to data mining bias, if many patterns are just chance results, as is surely the case with stock return data. To address this problem, the asset pricing literature recommends restricting the search to patterns consistent with theory (Cochrane (2005) and Harvey (2017)). However, recent empirical evidence finds this method is ineffective, even for theories published in top finance journals (Chen, Lopez-Lira, and Zimmermann (2022)).

We offer a different solution. Instead of mining data less, we recommend mining data *rigorously*. Rigorous data mining means conditioning interesting results on the fact that they come from searching through data. This conditioning can be achieved using empirical Bayes (Robbins (1956), Efron and Morris (1973), and Efron (2012)). Rigorous data mining also means that the search should be systematic, as is commonly done in high-throughput biology and chemistry (Yang et al. (2021)). Ironically, systematic search implies that asset pricing should involve *more* data mining, not less.

We use empirical Bayes (EB) to mine for out-of-sample returns among 136,000 long-short trading strategies. The trading strategies are constructed from systematically searching data on accounting ratios, past returns, and stock tickers. Through this "high-throughput asset pricing," we construct a portfolio with out-of-sample returns that are comparable to the returns from the best journals in finance.

Our data-mined portfolio is the simple average of the top 1% of strategies, based on EB-predicted Sharpe ratios. It earns out-of-sample returns of 5.7% per year over the 1983-2020 sample, compared to the mean return of 5.9% per year found by averaging the 200 published strategies from Chen and Zimmermann (2022). But unlike the published strategies, which were selected with knowledge of stock return patterns that occurred in the 1980s and 1990s, our strategies can be constructed using only information available in real time.

In fact, even naively mining for the largest Sharpe ratios leads to publication-like performance. We provide a theoretical explanation for this phenomenon in Proposition 1, which shows that under standard statistical practices (Fisher 1925), naive data mining often selects the same set of strategies as an ideal Bayesian. However, while the naively-selected strategies may be optimal, naive performance estimates are distorted, illustrating the importance of rigorously data mining with EB.

The top 1% portfolio selected by EB provides insights into the nature of return predictability. 91.0% of strategies in this portfolio are equal-weighted accounting ratio strategies. Almost all of the remainder are equal-weighted past-return strategies. Moreover, the returns of the top 1% strategy are concentrated in the pre-2004 data. These facts are consistent with the theory that predictability is largely due to limited attention and the slow incorporation of information into stock prices (Peng (2005); Chordia, Subrahmanyam, and Tong (2014)).

Other facts shed light on the drivers of the recent decline in cross-sectional predictability. We find that the returns of the top 5% and top 10% of portfolios are also concentrated in the pre-2004 data. These strategies are enormous in number: the top 5% consists of 6,305 strategies, and the top 10% consists of 12,610. As many of these strategies are unlikely to be found in academic journals, this suggests that the key driver of the recent declines in predictability is improvements in information technology (Chordia, Subrahmanyam, and Tong (2014)), rather than investors learning from academic publications (McLean and Pontiff (2016)). Consistent with this idea, we find that the top 20 strategies according to predicted Sharpe ratios using data available in 1993 have themes rarely seen in academic journals, like mortgage debt, growth in interest expense, and depreciation. Themes that were popular in academia in 1993, like book-to-market, momentum, and sales growth are missing from this list.

Overall, high-throughput asset pricing provides not only a method for dealing with look-ahead bias, but also a more rigorous method for documenting asset pricing facts. We post our strategy returns and code publicly, and encourage future researchers to use these methods.

Unlike many big data methods, EB provides a transparent intuition. In essence, EB measures the distance between the empirical t-stat distribution and the standard normal null. Ticker-based strategies have t-stats that are extremely close to the null, implying no predictability. In contrast, equal-weighted accounting t-stats are too fat tailed to be consistent with the null, implying strong predictability. Thus, just by visually inspecting the t-stat distributions, one can see where predictability is concentrated.

EB provides highly accurate predictions in pre-2004 data. We construct 120 portfolio tests using the 136,000 data-mined strategies, and compare EB-predicted returns with out-of-sample returns. In almost all of the 120 portfolios, the EB predictions are within 2 standard errors of the out-of-sample mean.

Post-2004, EB has more difficulty with accuracy, though it still captures broad patterns in out-of-sample returns. Compared to pre-2004, predicted returns are closer to zero, and only equal-weighted accounting strategies show notable predicted returns. However, out-of-sample returns are even closer to zero than predicted. This difficulty might be expected given the rise of information technology around 2004, which likely led to a structural break in predictability (Chordia, Subrahmanyam, and Tong 2014; Kim, Ivkovich, and Muravyev 2021). Our EB predictions are constructed using a simple 20-year rolling window, and thus fail to account for this break. This difficulty suggests that a smart data miner armed with theory might have understood the implications of the internet, and could perhaps have performed much better than our theory-free EB mining process.

We also illustrate how improper use of multiple testing statistics can lead to poor data mining results. We demonstrate this possibility using Harvey, Liu, and Zhu's (2016) recommended method for false discovery control. Harvey, Liu, and Zhu (2016) recommend applying Benjamini and Yekutieli's (2001) Theorem 1.3 to construct a t-stat hurdle that controls the false discovery rate (FDR) at the 1% level. Nearly all of our 136,000 trading strategies fail to meet this hurdle, suggesting that there are few interesting patterns in this data. But in fact, simple out-of-sample tests show there are thousands of strategies with notable out-of-sample returns. We find similar results following the recommended multiple testing control in Chordia, Goyal, and Saretto (2020), which is based on Romano and Wolf (2007). In contrast, the Storey (2002) FDR control recommended in Barras, Scaillet, and Wermers (2010), captures the majority of notable portfolios.

Fortunately, this error can be avoided by rigorously studying the statistics. According to Benjamini and Yekutieli (2001), their Theorem 1.3 is "very often unneeded, and yields too conservative of a procedure." This negative sentiment is echoed in Efron's (2012) textbook on large scale inference. In contrast, the EB methods we use are recommended for settings like ours in Chapter 1 of Efron (2012), as well as Chapters 6 and 7 of Efron and Hastie (2016).[1] The statistics literature has relatively little to say about the method recommended in Chordia, Goyal, and Saretto (2020). We provide our own characterization, which illustrates how this method is appropriate if selecting a null strategy is catastrophic. But using the standard null, that the mean long-short return (or alpha) is zero, Chordia

[1]A brief explanation of why Benjamini and Yekutieli (2001) Theorem 1.3 is excessively conservative is found in Section 2.5 of Chen (2024a).

et al.'s method implies unneeded conservatism.

## 1.1 Related Literature

We add to Yan and Zheng (2017) and Chen, Lopez-Lira, and Zimmermann (2022), who document that mining accounting data can produce substantial out-of-sample returns. Accounting data is important: Chen et al. find that mining ticker variables leads to out of sample returns of approximately zero. Thus, one needs a method for identifying the predictive power of accounting data in real time. Our empirical Bayes formulas provide one such method.

The literature on multiple testing in asset pricing features disagreement on both the methods that should be used and the empirical extent of multiple testing problems. Chen and Zimmermann (2020); Chen and Velikov (2022); and Jensen, Kelly, and Pedersen (2023) recommend empirical Bayes shrinkage. In contrast, Harvey, Liu, and Zhu (2016); Harvey and Liu (2020); and Chordia, Goyal, and Saretto (2020) recommend conservative false discovery controls, much more conservative than the FDR methods in Barras, Scaillet, and Wermers (2010). We show how empirical Bayes shrinkage and the recommended method from Barras, Scaillet, and Wermers (2010) leads to much more accurate inferences. More recently, Marrow and Nagel (2024) use empirical Bayes to study past return signals, with a focus on signal interactions and optimal weighting of more recent data.

In contrast to the intuition that simplicity is a virtue, we find that studying an enormous number of potential predictors leads to insights about the nature of return predictability. A similar theme is found in Kelly, Malamud, and Zhou (2024) and Didisheim et al. (2023), who illustrate the "virtue of complexity" in the modeling of expected returns.

# 2 Data and Methods

We describe the data (Section 2.1) and how we rigorously mine it (Sections 2.2-2.4).

## 2.1 Data on 136,000 Trading Strategies

Table 1 describes our data-mined strategies. The strategies are either based on accounting ratios, past returns, or tickers. Accounting ratio strategies are taken from Chen, Lopez-Lira, and Zimmermann (2022).[2] The past return and ticker strategies are inspired by Yan and Zheng (2017) and Harvey (2017), respectively, but we generate our own strategies in order to ensure that the number of strategies is comparable across data sources and to ensure that each type of strategy consists of many distinct strategies.[3]

[Table 1, Overview of Trading Strategies, about here]

A key feature of these strategies is that they are *not* selected based on having notable historical returns. Instead, they are constructed to systematically explore various types of data. So unlike most datasets in asset pricing (e.g. Ken French's size- and B/M-sorted portfolios; Chen and Zimmermann (2022)), ours is arguably free of data mining bias. Indeed, Table 1 shows that the median sample mean return is close to zero for all sets of strategies.

In high-throughput research, the median measurement is relatively unimportant. What matters is that the extreme measurements show promise for, say, a pharmaceutical intervention or cancer prediction. The extreme measurements in Table 1 suggest that accounting and past return data show promise for predicting returns. These data lead to mean returns that can exceed 5 percent per year in absolute value.

For further details on the strategy definitions, see Appendix A or our github site.

## 2.2 Empirical Bayes Overview

The 136,000 strategies in Table 1 contain the potential for significant data mining bias. To understand the bias, let $r_i$ be a performance measure for strategy $i$ (e.g. mean return, alpha) and decompose it as follows:

$$r_i = \mu_i + \varepsilon_i \tag{1}$$

[2] We are grateful that the authors make their data publicly available.

[3] Results that mine data following Yan and Zheng (2017) and Harvey (2017) are similar and can be found in the first draft of our paper on arxiv.org or via our github site.

where $\mu_i$ is the actual performance and $\varepsilon_i$ is sampling error or luck.

Data mining involves selecting $i$ with large $r_i$. Suppose we set $\bar{r} \gg 0$, and search for $i^* \in \{1;2;\ldots;136,000\}$ such that $r_{i^*} = \bar{r}$. This practice is dangerous because one might think $\bar{r}$ is a good estimate of $\mu_{i^*}$. However, $\bar{r}$ is in fact biased upward

$$\begin{aligned}\bar{r} &= E\left(r_{i^*}|r_{i^*} = \bar{r}\right) \\ &= E\left(\mu_{i^*}|r_{i^*} = \bar{r}\right) + \underbrace{E\left(\varepsilon_{i^*}|r_{i^*} = \bar{r}\right)}_{>0} > \mu_{i^*}.\end{aligned} \tag{2}$$

Selecting for large $r_i$ also selects for large $\varepsilon_i$, leading to $E\left(\varepsilon_{i^*}|r_{i^*} = \bar{r}\right) > 0$ and the bias in Equation (2).

To data mine safely, one needs to remove the luck term $E\left(\varepsilon_i|r_{i^*} = \bar{r}\right)$. This term is just a conditional expectation, so it can be computed using Bayes rule, provided one has a probability model for $\mu_i$ and $r_i$.

Suppose one has a probability model, with parameter vector $\Omega$. The bias can then be removed by computing

$$E\left(\mu_{i^*}|r_{i^*} = \bar{r};\hat{\Omega}\right) = \bar{r} - E\left(\varepsilon_{i^*}|r_{i^*} = \bar{r};\hat{\Omega}\right) \tag{3}$$

where $\hat{\Omega}$ is a consistent (frequentist) estimate of the probability model parameters. This method, of applying frequentist estimates to Bayesian formulas is known as "empirical Bayes" (Robbins (1956) and Efron and Morris (1973)).

Equation (3) conditions on only one statistic regarding strategy $i$. A more optimal estimate uses more information

$$E\left(\mu_{i^*}|r_{i^*} = \bar{r}, X_i = \bar{X};\hat{\Omega}\right) = \bar{r} - E\left(\varepsilon_{i^*}|r_{i^*} = \bar{r}, X_i = \bar{X};\hat{\Omega}\right) \tag{4}$$

where $X_i$ is a vector of additional statistics for strategy $i$ and $\bar{X}$ is a realized value of $X_i$. For example, $X_i$ can include the standard error of $r_i$, the portfolio weighting (equal- or value-weighted), and the signal data source (accounting, past returns, tickers).

We use Equation (4) to search our 136,000 strategies for large expected returns. We will not use economic theory to determine the probability model, and thus our search is largely atheoretical. However, we recognize the bias that comes from such a search (Equation (2)), and carefully correct for it. Thus, we describe our methods as "rigorous data mining."

## 2.3 Optimal Naive Data Mining

In empirical asset pricing, we are often interested in two questions:

1. What are the best strategies?

2. What is the performance of the best strategies?

If one is interested *only* in the first question, then there is a sense in which naively mining data, without accounting for data mining bias, is often optimal.

To understand this, we add structure to the model. First, explicitly define the additional statistics $X_i$:

$$X_i = [D_i, \mathrm{SE}_i] \tag{5}$$

where $D_i$ is the strategy "family" (e.g. equal-weighted accounting) and $\mathrm{SE}_i$ is the standard error of $r_i$. Actual performance follows

$$\mu_i | X_i \sim g_{D_i,\mathrm{SE}_i}(\cdot) \tag{6}$$

where $g_{D_i,\mathrm{SE}_i}(\cdot)$ is a distribution that depends on $D_i$ and $\mathrm{SE}_i$. Measured performance follows

$$r_i | \mu_i, X_i \sim f_{\mu_i,\mathrm{SE}_i}(\cdot) \tag{7}$$

where $f_{\mu_i,\mathrm{SE}_i}(\cdot)$ is a distribution that depends on $\mu_i$ and $\mathrm{SE}_i$. This is a hierarchical structure, where the strategy family determines the actual performance, which in turn determines the measured performance.

Second, define data mining. Naive data mining chooses a hurdle $h$ and then selects strategies

$$\{i : r_i > h\} \tag{8}$$

In contrast, EB data mining uses the bias-adjusted measure to select strategies

$$\{i : E\left(\mu_i \mid r_i, X_i\right) > h'\} \tag{9}$$

where $h'$ is chosen to select the same number of strategies as in naive data mining.

In general, Equations (8) and (9) imply different sets of strategies. However, under some natural conditions, the selections are identical:

**Proposition 1.** *Consider the following two conditions:*

1. *The performance measure satisfies*

$$r_i \mid \mu_i, X_i \sim Normal\left(\mu_i, SE^2\right) \tag{10}$$

*where SE is a constant.*

2. *The hurdle h satisfies*

$$\Pr\left(r_i > h | D_i\right) = 0 \quad if\, D_i \in \mathcal{D} \tag{11}$$

$$\mu_i | X_i \sim g_{SE_i}(\cdot) \quad if\, D_i \in \mathcal{D} \tag{12}$$

*where $\mathcal{D}$ is a subset of the possible strategy families and $g_{SE_i}(\cdot)$ is distribution with positive variance that does not depend on $D_i$.*

*If conditions 1 and 2 hold, then naive data mining selects the same set of strategies as empirical Bayes.*

Conditions 1 and 2 arise naturally when using long samples (e.g. 300 months of returns), standardized performance measures (e.g. t-statistics), and strict statistical hurdles (e.g. 5% critical levels). Under these conditions, actual performance is a strictly increasing function of only the measured performance, as proved in Appendix B.1. As a result, data-mined performance provides a reliable signal of actual performance, even if the magnitudes are distorted. The proposition assumes some exact conditions, and leading to identical selections, but approximate conditions would likely lead to similar selections.

One interpretation of Proposition 1 is that Fisher's (1925) focus on t-statistics set future researchers up for success, even in the modern era of big data.

On the other hand, Fisher would likely have been unsatisfied with finding the best strategies. He most likely would implore us to find unbiased estimates for these best performers. Thus to rigorously mine data, one should still apply empirical Bayes.

## 2.4 Empirical Bayes Implementation

We select as our performance measure the t-statistic on the raw long-short return, and assume that standard errors are precisely measured, implying

$$r_i|\mu_i, X_i \sim \text{Normal}\left(\mu_i, 1\right) \tag{13}$$

The latent performance is a mixture of two normals that depends on the strategy family $D_i$.

$$\mu_i \mid (X_i, D_i = d) \sim \begin{cases} \text{Normal}\left(\theta_{d,1}, \sigma^2_{d,1}\right) & \text{with prob } \lambda_d \\ \text{Normal}\left(\theta_{d,2}, \sigma^2_{d,2}\right) & \text{otherwise} \end{cases}. \tag{14}$$

where $d$ is one of the six strategy families that comes from combining three data sources (accounting, past returns, tickers) with two portfolio formation methods (equal-weighted and value-weighted). Mixture normals are parsimonious, easy to understand, and yet allow for skewness and fat tails.

We then estimate $\Omega \equiv \left[\theta_{d,1}, \sigma^2_{d,1}, \theta_{d,2}, \sigma^2_{d,2}, \lambda_d\right]_{d=1,\ldots,6}$ using quasi-maximum likelihood. The quasi-likelihood is computed using the distr package (Ruckdeschel et al. (2006)). Optimization of $\Omega$ uses nloptr (Johnson (2007)).This estimation is done using the past 20 years of long-short returns, separately for each "forecasting year" spanning 1983-2019.

Finally, we recover EB predictions by computing Equation (4) with distr, which produces an EB prediction of the expected return in units of standard errors. The EB predicted return is just Equation (4) multiplied by the standard error. Similarly, the EB predicted Sharpe ratio is Equation (4) multiplied by the square root of the number of periods in the sample.

For further details see Appendix B or our github site.

# 3 Performance of the Best Data Mined Strategies

We show that data mining leads to research-like out-of-sample returns (Section 3.1) and take a look at which kinds of strategies are identified by data mining (Section 3.2). We also provide intuition for why data mining produces such high returns (Section 3.3).

## 3.1 Out-of-Sample Returns

Can data mining generate out-of-sample returns? To answer this question we construct simple out-of-sample portfolio tests.

Each year, we sign strategies to have positive predicted returns, and then form portfolios that equally-weight strategies in the top $X$% of predicted Sharpe ratios. We use both EB predictions and standard naive predictions. We examine $X$ = 1, 5, and 10. For comparison, we also examine a portfolio that equally-weighs published strategies from the Chen and Zimmermann (2022) dataset.

Table 2 shows the result. Using empirical Bayes (EB Mining), the top 1% of strategies perform similarly to strategies published in top finance journals. Over the full 1983-2020 sample, the top 1% portfolio earns 5.70% per year, compared to the 5.88% return from published strategies. The Sharpe ratio from EB mining is smaller, at 1.46 vs 2.03 for published strategies. However, unlike the EB-mined strategies, which are formed using only information available in real-time, the published strategies contain look-ahead bias. Indeed, if we focus on strategies in top journals that were published pre-2004, the performance is very similar to the EB-mined strategies in terms of either mean returns or Sharpe ratios.

[Table 2, Returns of Long-Short Portfolios Data-Mined, about here]

The EB-mined returns are robust. The top 5% and top 10% of data-mined strategies also perform well and are extremely statistically significant, indicating that the performance of the top 1% is not driven by outliers.

Panel B shows that even naive data mining produces research-like returns. Simply choosing the top 1% of strategies based on their past Sharpe ratios leads to an out-of-sample Sharpe ratio of 1.45. This is almost exactly the same as the Sharpe ratio from the top 1% using EB mining, consistent with Proposition 1. The top 5% and top 10% of naive strategies underperform a bit relative to EB mining, but the intuition behind Proposition 1 still goes through.

Figure 1 takes a closer look by plotting the value of $1 invested in each portfolio over time. The top 1% data-mined strategies have similar performance to published strategies throughout the figure. All portfolios show relatively little cyclicality during the recessions of 1991, 2009, and 2020. Indeed, the returns are fairly consistent throughout the chart, with an important caveat.

[Figure 1, Cumulative Long-Short Returns, about here]

The caveat is that returns are concentrated in the pre-2004 sample. This is seen in the flattening of the solid line in Figure 1 around 2004. The EB Mining top 1% portfolio returns 8.17% per year from 1983-2005, compared to just 2.03% from 2005-2020. A similar decay is seen across all portfolios, both data-mined and academic. This decay is consistent with Chordia, Subrahmanyam, and Tong (2014) and Chen and Velikov (2022), who argue that the rise of information technology reduced return predictability.

Overall, we find that one can find long-short returns comparable to those from the best journals in finance, just by mining data, with little thought about the underlying economics. Moreover, rigorous data mining can discriminate between data sources that have no information about future returns, like stock market tickers, from data that is rich in information, like accounting ratios. Unlike the published strategy returns, our returns can be found using only information available in real-time. These results show that high-throughput methods provide a bias-free approach to studying stock market predictability. Our strategy returns and code are public, and we encourage future researchers to use these methods.

## 3.2 The Composition of the Top 1%

Table 3 takes a closer look at the top 1% strategies produced by rigorous data mining. Panel A shows that 91.0% of the top 1% come from the equal-weighted accounting family and 8.6% come from equal-weighted past returns. The other strategy families comprise a negligible part of the top 1%. Ticker strategies are completely absent.

[Table 3, Description of Top 1% Data-Mined Strategies, about here]

Taken with Table 2, these results show that cross-sectional predictability is concentrated in accounting data, small stocks, and pre-2004 samples. These stylized facts offer a parsimonious description of the "factor zoo." Theories that wish to capture the big picture of cross-sectional predictability should be consistent with these facts. For example, slow diffusion of economic information is consistent, as this diffusion would be especially slow in small stocks and before the internet era. In this way, high throughput asset pricing provides a way to not only identify out-of-sample returns, but to also provide insight into the underlying economics.

Panel B of Table 3 shows that many of the top 1% strategies are quite far from the predictors noted in the academic literature. In 1993, academics were focused on predictors like book-to-market, 12-month momentum, and sales growth (Fama and French (1992); Jegadeesh and Titman (1993); Lakonishok, Shleifer, and Vishny (1994)). None of these predictors are in the top 20 strategies based on predicted Sharpe ratios from rigorous data mining. Instead, the common themes from data mining include shorting stocks with high or growing debt, as well as buying stocks with high depreciation, depletion, and amortization. Another theme is buying stocks with high returns in quarters $t$ minus 17 and 18.

Based on textbook risk-based or behavioral asset pricing, one might expect that these data-mined predictors will average zero returns out-of-sample. But this is not the case. The realized Sharpe ratios for these strategies in the 10 years after 1993 averages around 1.0 ("SR OOS" column).

Panel B of Table 3 focuses on 1993 because well-known predictability papers were published around that time (e.g. Fama and French (1993)). In other years, the top 20 list is different, though shorting variables related to debt growth remains a common theme. For further details see Appendix Tables A.1 and A.2.

## 3.3 Shrinkage Intuition

Unlike many big data and machine learning methods, empirical Bayes has a transparent intuition. The intuition can be seen in a special case of the prediction Equation (4). If $\mu_i \mid (X_i, D_i = d) \sim \text{Normal}\left(0, \sigma_d^2\right)$, we have

$$E\left(\mu_i \mid r_i = \bar{r}, X_i = \bar{X}, D_i = d\right) = \left[1 - \frac{1}{\widehat{\text{Var}}\,(r_i | D_i = d)}\right] \bar{r}, \tag{15}$$

where $\widehat{\text{Var}}\,(r_i | D_i = d)$ is an estimate of the cross-strategy variance of performance measures among strategies with data family $d$.

This expression says that rigorous mining involves shrinking performance measures $r_i$ toward zero at a rate of $\frac{1}{\widehat{\text{Var}}(r_i|D_i=d)}$. $\widehat{\text{Var}}\,(r_i|D_i = d)$ measures how far the data are from the null of $r_i \sim \text{Normal}\,(0,1)$, which we imposed in Equation (13). If there is no predictability, then $r_i \sim \text{Normal}\,(0,1)$, $\widehat{\text{Var}}\,(r_i|D_i = d) \approx 1$, and all $r_i$ are shrunk to zero. But if data are far from the null, then a large $r_i$ is a signal of large $\mu_i$—even if $r_i$ is found from searching tens of thousands of strategies, unguided by economic theory.

Figure 2 shows that equal-weighted accounting strategies (upper left) are far from the null using data from 1964 to 1983. Equal-weighted past return strategies (middle left) also show a notable deviation. In contrast, the other strategy families are quite close to the null. Indeed, for both families of ticker-based strategies, the null is a very good fit for the data.

[Figure 2, Distribution of t-stats in 1983, about here]

Accordingly, Equation (15) implies that the strategies with strong actual performance will be found in equal-weighted accounting and equal-weighted past-return strategies. This intuition is consistent with Panel A of Table 3, which shows that the vast majority of the best data-mined strategies come from these families.

Compared to data available in 1983, all strategy families are closer to the null using data from 1985-2004, as seen in Figure 3. All value-weighted families are very close to the null, implying that predictability in large stocks is essentially gone. The long left tail in equal-weighted past return strategies also disappears. Only equal-weighted accounting strategies are visually far from the null. These results imply that predictability is concentrated in the earlier part of the sample.

[Figure 3, Distribution of t-stats in 2004, about here]

The intuition in Figures 2 and 3 is so simple that one might even skip the quasi-maximum likelihood estimation. Just looking at these charts, and the distance between the data and the null, one can already tell that predictability is concentrated in small stocks, accounting data, and the earlier sample. That is, one can already tell where predictability is concentrated, if one understands the intuition in Equation (15).

# 4 Empirical Bayes Prediction Accuracy Across the Cross-Section

This section takes a closer look at the EB predictions and accuracy. We see when and where EB predictions are successful and when they struggle.

## 4.1 EB Prediction Accuracy 1983-2004

To examine accuracy, we use out-of-sample portfolio sorts. For each year and each strategy family, we form 20 portfolios by sorting strategies into equal-sized groups based on the past 20 years of mean returns. We then predict the mean returns for each portfolio by averaging the EB predictions (Equation (4)), which are also based on the past 20 years of data. Finally, we form a portfolio that equally-weighs strategies in each group and hold for one year (the "out-of-sample" periods).

Figure 4 shows the in-sample, predicted, and out-of-sample returns for each portfolio, averaged over the out-of-sample periods from 1983 to 2004. For all six families, there are sizable in-sample returns (dashed line) in the extreme in-sample groups. For accounting strategies, in-sample returns are as extreme as -11% per year. A naive read of this result is that one can flip the long and short legs and find +11% returns out-of-sample. Past return strategies see a similar $\pm 10$ percent return in the extreme groups. Even ticker-based strategies show in-sample long-short returns of up to 4 percent per year.

[Figure 4, Empirical Bayes Predictions 1983-2004, about here]

However, the predicted returns are typically much closer to zero. In fact, for both ticker-based strategy families, the predicted return (solid line) is almost exactly zero for all 40 in-sample groups. This result is intuitive given how close the ticker t-stats are to the null of no predictability (Figure 3). This closeness implies that the extreme returns can be entirely accounted for by luck, and so shrinkage should be 100% (Equation (15)). Significant shrinkage is also seen in value-weighted accounting strategies (top right panel). Rigorous data mining recommends that the extreme returns of around -8% and +9% (dashed line) be shrunk down to about -3% and +2 (solid line), respectively.

Rigorous mining predicts much higher returns in equal-weighted accounting strategies (upper left panel). For these strategies, the predicted returns are actually not far from the in-sample return. This result is consistent with Chen and Zimmermann (2020), who find shrinkage of only 12% for published anomalies, which are largely equal-weighted and based on accounting variables. Predictability is also seen in both families of past return strategies.

These predictions are borne out in out-of-sample returns (markers with error bars). The first group of EW accounting strategies returns -8 percent per year out-of-sample from 1983-2004, almost exactly the same as the EB prediction. Similar accuracy is seen throughout all 120 bins in Figure 4.

These results show that rigorous data mining offers economic insights that are difficult to derive from theory. While theories of slow information diffusion may tell you that predictability is concentrated in small stocks, accounting signals, and pre-2004 data, they are unlikely tell you how much predictability there is. In contrast, empirical Bayes provides quantitative, accurate estimates of the precise amount of predictability.

## 4.2 EB Prediction Accuracy 2004-2020

We split our OOS tests in the mid-2000s, motivated by the idea that there was likely a structural break during this period due to the rise of information technology (Chordia, Subrahmanyam, and Tong (2014)). Comparing the distribution of t-stats available in 1983 vs 2004 supports the idea that the structure of financial markets changed (see Section 3.3).

[Figure 5, Empirical Bayes Predictions 2004-2020, about here]

This structural change can be seen by comparing Figure 5 (EB predictions 2004-2020) to Figure 4 (EB predictions 1983-2004). In all panels, the predicted returns shift closer to zero post-2004. Most notably, the predictability that was present in past return strategies pre-2004 is largely gone. Consistent with these predictions, the past return portfolios show a flat or even negative relationship between out-of-sample and in-sample returns post-2004. A similar weakening of EB predictions and flattening of out-of-sample returns is seen in the accounting VW family.

An exception to this pattern is the family of equal-weighted accounting ratio strategies (top left). In this chart, the shrinkage is still relatively small, with EB predictions implying returns as extreme as -9 percent per year. This prediction and others in this panel miss the mark: the out-of-sample returns are much closer to zero throughout this panel.

This poor accuracy is natural given the fact that the estimations use a rolling window consisting of the past 20 years of data. This fixed window implies that, for much of the period 2004-2020, our estimates rely on data from a time when accounting statements needed to be retrieved by traditional (snail) mail for investors without special access to the SEC reading room (Bowles et al. (2023)).

This result implies an important role for economic theory: when structural breaks occur, there is no way for data mining to provide a clear understanding of the economy, no matter how rigorously the mining is done. Theory is sometimes used this way in economics and finance, but this is typically not the case. Instead, theory is typically used to understand patterns found in long samples of data, spanning many decades. In our view, the future of theory is bright for theorists who study structural breaks, even in the era of big data. Indeed, a smart data miner armed with theory might have understood the implications of the internet for stock return predictability, and could perhaps have performed much better than our theory-free EB mining process.

# 5 Comparison with False Discovery Controls

Our main analysis corrects for data mining bias using empirical Bayes shrinkage, following Chen and Zimmermann (2020); Chen and Velikov (2022); and Jensen, Kelly, and Pedersen (2023). An alternative approach is to use false discovery controls, following Harvey, Liu, and Zhu (2016); Barras, Scaillet, and Wermers (2010); or Chordia, Goyal, and Saretto (2020). The ideal approach remains an unsettled question. Our dataset of 136,000 trading strategies provides a natural testing ground.

We examine the following false discovery controls:

1. **BY1.3 (1%)**: Harvey, Liu, and Zhu (2016) (HLZ) recommend using Benjamini and Yekutieli's (2001) Theorem 1.3 at the 1% level. HLZ is likely the most influential paper on multiple testing in empirical asset pricing.

2. **Storey (10%)**: Barras, Scaillet, and Wermers (2010), which introduced false discovery methods to finance, study the Storey (2002) algorithm at the 10% level.

3. **RW (5%, 5%)**: Chordia, Goyal, and Saretto (2020) recommend combining Romano and Wolf's (2007) Algorithms 4.1 and 2.1. These algorithms require two parameters, both of which Chordia et al. set to 5%.

For each year and each strategy family, we apply these methods using the past 20 years of data to estimate a t-statistic hurdle. We then examine whether these hurdles are able to separate strategies with high out-of-sample returns from those with low out-of-sample returns. This structure is the same as in Section 4.

Figure 6 shows the results. The vertical lines show the mean hurdle across all years. The markers show the mean out-of-sample returns of portfolios formed by equally weighting strategies, sorted into 20 groups based on the in-sample t-statistic. Groups of strategies that a false discovery control declares "significant" lie on the outside of the respective vertical lines.

[Figure 6, False Discovery Controls, about here]

The BY1.3 (1%) and RW (5%, 5%) methods miss out on the majority of portfolios with notable out-of-sample performance. Out of the 5 groups that have out-of-sample returns of at least 3% per year, only 1 lies outside of the solid lines corresponding to BY1.3 (1%). Only 3 of 5 lie outside the dot-dashed lines corresponding to RW (5%, 5%). The dashed line, corresponding to Storey (10%), performs much better, capturing 4 of the 5 portfolios. Similar results are found using alternative parameter choices examined by HLZ; Harvey and Liu (2020); and Barras, Scaillet, and Wermers (2010) (see Appendix Figure A.1).

Thus, Storey (10%) provides an easy-to-compute alternative to empirical Bayes. However, Storey cannot provide bias-adjusted performance estimates that are naturally available from empirical Bayes. Overall, our results imply that Storey forms a strong first step for rigorous data mining, while empirical Bayes is recommended for more refined estimates. These results are broadly consistent with the statistics literature, which generally recommends Storey as a preliminary examination, while suggesting empirical Bayes for greater precision (e.g. Benjamini 2010; Efron 2012).

We discuss this literature and the algorithms in more detail below.

## 5.1 Benjamini and Yekutieli (2001) Theorem 1.3

Harvey, Liu, and Zhu (2016) (HLZ) recommend using Benjamini and Yekutieli's (2001) Theorem 1.3. Several followups to the influential HLZ paper use this method, including Harvey and Liu (2020) and Chordia, Goyal, and Saretto (2020); and Jensen, Kelly, and Pedersen (2023).

We state the theorem number 1.3 because the bulk of the original paper focuses on Theorem 1.2. Indeed, Benjamini and Yekutieli (2001) describe Theorem 1.3 as "very often unneeded, and yields too conservative of a procedure" (page 1183). In his textbook on large scale inference, Efron (2012) agrees, stating that the theorem represents a "severe penalty" and is "not really necessary" (section 4.2). Moreover, the statistics literature uses the "BY algorithm" to refer to Benjamini and Yekutieli (2005), which is an entirely different procedure (e.g. Efron 2012 Chapter 11.4).

BY1.3 begins by choosing a parameter $q^*$ and then solving

$$h_{\text{HLZ},q^*} \equiv \min_{h>0} \left\{ h : \left[ \frac{\Pr(|Z| > h)}{\text{Share of } |t_i| > h} \right] \pi_{\text{BY1.3}} \le q^* \right\} \tag{16}$$

where $t_i$ is the t-statistic for strategy $i$, $Z$ is a standard normal random variable,

$$\pi_{\text{BY1.3}} \equiv \sum_{i=1}^{N} \frac{1}{i}, \tag{17}$$

and $N$ is the number of strategies in the year-family. Benjamini and Yekutieli's (2001) Theorem 1.3 proves that this algorithm implies a false discovery rate $\le q^*$. BY1.3 amounts to modifying the seminal Benjamini and Hochberg (1995) algorithm with a constant factor, $\pi_{\text{BY1.3}}$. This modification makes the algorithm more conservative.

HLZ recommend this conservative approach, claiming Benjamini and Hochberg (1995) "is only valid when the test statistics are independent or positively dependent" (page 21). This statement is false. Storey and Tibshirani (2001) and Storey, Taylor, and Siegmund (2004) show validity under weak dependence assumptions (see also Chen 2024b).

HLZ are also conservative in their choice of $q^*$. For their main results, they use $q^* = 1\%$ citing the fact that the "significance level is subjective," though they also examine $q^* = 5\%$ for robustness. In contrast, the statistics literature generally recommends $q^* = 5\%$ or 10% (e.g. Benjamini 2010; Efron 2012).

Given this context, it is perhaps unsurprising that BY1.3 (1%) fails to identify most out-of-sample performers in Figure 6. BY1.3 (5%) performs somewhat better, identifying 2 out of 5 groups with out-of-sample returns of at least 3% per year (Appendix Figure A.1).

## 5.2 Storey's (2002) FDR Control

While HLZ recommend modifying Benjamini and Hochberg (1995) to be more conservative, much of the statistics literature goes in the opposite direction, modifying Benjamini and Hochberg (1995) to be more aggressive. In finance, Barras, Scaillet, and Wermers (2010) take this approach.

Barras et al. recommend the Storey (2002) algorithm, which can be written as

$$h_{\text{Storey},q^*} \equiv \min_{h>0} \left\{ h : \left[ \frac{\Pr(|Z| > h)}{\text{Share of } |t_i| > h} \right] \pi_{\text{Storey}} \leq q^* \right\} \tag{18}$$

where $t_i$ is the t-statistic for strategy $i$, $Z$ is a standard normal random variable,

$$\pi_{\text{Storey}} = \frac{\text{Share of } |t_i| \leq 1.0}{\Pr(|Z| \leq 1.0)} = \frac{\text{Share of } |t_i| \leq 1.0}{0.68} \tag{19}$$

and the cutoff of 1.0 is selected for ease of interpretation. Storey (2002) proves that this algorithm implies a false discovery rate $\leq q^*$ under independence assumptions, though Storey and Tibshirani (2001) and Storey, Taylor, and Siegmund (2004) extend this result to weak dependence.

Comparing Equations (18)-(19) to the Equations (16)-(17), we see that the only difference is the constant factor, $\pi_{\text{BY1.3}}$ vs $\pi_{\text{Storey}}$. These constants are qualtitatively different: $\pi_{\text{BY1.3}} = \sum_{i=1}^{N} \frac{1}{i} \approx 0.6 + \log N \gg 1$ , while $\pi_{\text{Storey}} \leq 1.0$. As shown in Storey (2002), $\pi_{\text{Storey}}$ can be interpreted as an estimate of the probability that a strategy is null, which can be at most 1.0. Other statistics papers that recommend a constant that is at most 1.0 include Benjamini and Hochberg (2000), Efron, Tibshirani, et al. (2001), Genovese, Roeder, and Wasserman (2006), and Benjamini, Krieger, and Yekutieli (2006).

Barras, Scaillet, and Wermers (2010) do not emphasize a particular choice of $q^*$, and instead examine values ranging from 5% to 20%. Figure 6 uses $q^* = 10\%$, because Barras et al. use 10% in their illustrative examples.

Once again, given the support from the statistics literature, it is perhaps unsurprising that Storey (10%) and (20%) perform well. Equations (18)-(19) are easy to implement, making it a useful alternative to our empirical Bayes method.

However, there are two downsides to using Storey. A simple, symmetric testing algorithm like Equations (18)-(19) does not handle skewed distributions well. This limitation may explain why Storey struggles to identify out-of-sample performers in past return strategies, which feature a long right tail (Figure 2). The second is that Storey cannot provide bias-adjusted performance estimates. Such estimates are naturally available from a more general empirical Bayes method, and would provide clean connections with portfolio choice and asset pricing questions.

## 5.3 Romano and Wolf's (2007) FDP Risk Control

Chordia, Goyal, and Saretto (2020) recommend combining Romano and Wolf's (2007) Algorithms 4.1 and 2.1, which we refer to as "RW." This algorithm is a natural choice for asset pricing researchers, as its predecessor Romano and Wolf (2005) is motivated by data mining for CAPM anomalies. Like HLZ's method, the Romano and Wolf methods have been used in influential asset pricing papers, including Chordia, Goyal, and Saretto (2020); Engelberg et al. (2023); Heath et al. (2023); and Bodt, Eckbo, and Roll (2025).

Unlike Storey and BY1.3, the statistics literature has relatively little discussion of the Romano and Wolf methods. Neither Romano and Wolf (2005) nor Romano and Wolf (2007) is found in the textbooks Efron (2012) and Efron and Hastie (2016). The two papers are also not found in the review articles on multiple testing Benjamini (2010) and Benjamini (2020). Thus, we provide some discussion here.

The goal of RW can be written as follows: find an $h$ that ensures

$$\Pr(\text{FDP} > p^*) \leq q^* \tag{20}$$

where

$$\text{FDP} \equiv \frac{\text{Number of null strategies with } |t_i| > h}{\text{Number of strategies with } |t_i| > h} \tag{21}$$

and $p^*$ and $q^*$ are thresholds selected by the researcher. Null strategies are, typically, those with an actual performance of zero.

Figure 7 illustrates Equation (20), by simulating one of our QML estimates many times. We run 2,000 simulations, each one consisting of 29,000 strategies. For simplicity, we assume all strategiesare independent. The plot shows histograms of actual performance ($\mu_i$ in Equation (1)) for strategies that meet the hurdle $|t_i| > 3.0$, where $h = 3.0$ is selected for illustrative purposes. Using this chart, we can ask whether this $h = 3.0$ hurdle achieves Equation (20), and thus understand FDP risk control.

[Figure 7, FDP Risk Control Illustration, about here]

Since the $h = 3.0$ hurdle is quite stringent, the vast majority of strategies are non-null. However, there is still a risk that a strategy with $|t_i| > 3.0$ has near-zero actual performance, as seen in the left tail of the histogram. The FDP characterizes this risk. It is, approximately, the share of strategies in the first bin.[4] On average, the share of strategies in this bin is about 5% (bars), indicating that the FDR is approximately controlled at a 5% level.

Even though the FDP is on average about 5%, there is a risk that it is higher. This risk is seen in the lines of Figure 7, which plot extreme order statistics across the 2,000 simulations. The 95th percentile line implies the FDP exceeds 7% in 5% of simulations. To achieve FDP risk control with $p^* = 5\%$ and $q^* = 5\%$, a $h > 3.0$ is required. The RW method finds this $h$.

Thus, the RW method aims to control the tail risk of a tail risk. Such an algorithm is a natural choice if selecting a null strategy is catastrophic. In such a case, one may want to ensure not only that a null is highly improbable, but that the probability that a null is somewhat probable is also improbable. However, in the standard setting where the null is that the strategy has zero long-short return or zero alpha, then the RW method tends to imply extreme conservatism.

This conservatism leads to the results in Figure 6 and Appendix Figure A.1. Choosing $p^* = 0.05$ and $q^* = 0.05$ or 0.10, as in Chordia, Goyal, and Saretto (2020) and Harvey, Liu, and Saretto (2020), leads to hurdles that many notable out-of-sample performers fail to clear.

The RW method is rather complex. It uses cluster bootstrap methods, involves testing all possible subsets of selected sets of strategies, iterating over many possible tests and sets. We describe our implementation in Appendix B.3 and provide code in our Github repo.

# 6 Conclusion

We show that a solution to data mining bias is to mine data rigorously. We systematically search 136,000 long-short strategies and find out-of-sample performance comparable to academic research. Simply searching for strategies with the largest t-stats leads to publication-like out-of-sample performance, a fact we explain in a Bayesian model. While naive data mining leads to distorted performance estimates, empirical Bayes provides unbiased predictions in samples without structural breaks. The forecast errors around structural breaks suggest a

[4]More formally, one can consider the first bin to be an upper bound on the FDP (see Chen 2024b).

role for theory in the era of big data.

This high-throughput method shows that returns are concentrated in accounting signals, small stocks, and pre-2004 periods, consistent with mispricing and slow information diffusion theories. While these results could potentially be gleaned from a deep read of the anomalies literature, our method provides a scientific method for documenting these stylized facts. We provide our data and code publicly, and hope others follow in using high-throughput methods.

Our out-of-sample tests offer an intuitive method for comparing multiple testing methods. We find that methods popular in finance would lead researchers to miss out on the majority of signals with notable out-of-sample performance. In contrast, methods recommended by the statistics literature perform well.

# A Data Handling Details

## A.1 60,000 Accounting Ratio Strategies

We examine 60,000 accounting ratio strategies constructed by Chen, Lopez-Lira, and Zimmermann (2022). Inspired by Yan and Zheng (2017), Chen et al. construct 30,000 accounting ratio signals as follows. Let $X$ be one of 240 accounting variables from Compustat (+ CRSP market equity) and $Y$ be one of 65 accounting of these 240 variables that is positive for at least 25% of firms in 1963. Apply two transformations: $X/Y$ and $\Delta X/\text{lag}Y$ to get $240 \times 65 \times 2 \approx 30,000$ signals. Then form equal-weighted and value-weighted long-short decile strategies, leading to 60,000 strategies.

These strategies are downloaded from Andrew Chen's website. We are grateful to the others for making their data public.

## A.2 38,000 Past Return-Based Strategies

Inspired by Yan and Zheng (2017) , we construct past-return strategies as follows: Choose 4 quarters out of the past 20 quarters. Compute the first four central moments using the returns in these quarters. This leads to $\binom{20}{4} \times 4 = 19,380$ signals.

Add to this the return over any of the past 20 quarters, as well as the mean return over the past 2 and past 3 quarters. This adds $20 + 2$ signals, for a total of $19,380 + 22 = 19,402$ signals.

Finally, form equal-weighted and value-weighted long-short decile strategies.

We chose this approach, rather than the approach in Yan and Zheng (2017) for three reasons. The first is that we want to have a strategy list that is comparable in length to the length of our accounting ratio strategies. Yan and Zheng's method leads to "only" 4,080 signals. The second is that, while Yan and Zheng's methods are inspired by momentum and short-run reversal, we want to ensure that our methods do not incorporate knowledge that would come from reading finance publications. Last, we chose to reduce the amount of overlap across the different signals, which should lead to better properties of our EB estimator.

Earlier versions of our paper use Yan and Zheng's method and found similar results. These results can be found at arxiv.org.

### A.3 38,000 Ticker-Based Strategies

Inspired by Harvey (2017), we sort stocks into 20 groups based on the alphabetical order of the first ticker symbol. We then long any two of those groups and short two. Repeat using the 2nd, 3rd, and 4th ticker symbols. This yields $\binom{20}{4} \times 4 = 19,380$ long-short portfolios.

We chose not to follow Harvey (2017)'s approach in order to have a similar number of strategies as our accounting-based strategies. Harvey's method leads to "only" 6,000 ticker-based strategies.

Earlier versions of our paper used Harvey's method and found similar results. These results can be found at arxiv.org.

# B Theory and Estimation Details

## B.1 Proof of Proposition 1

*Proof.* Naive and EB mining will select the same strategies as long as, for $r_i > h$, $E(\mu_i|r_i, \text{SE}_i, D_i)$ does not depend on $\text{SE}_i$ or $D_i$ and is strictly increasing in $r_i$.

Condition 1 implies that $\text{SE}_i$ is constant. Thus $E(\mu_i|r_i, \text{SE}_i, D_i)$ does not depend on $\text{SE}_i$.

Condition 2 implies that, for $r_i > h$, $E(\mu_i|r_i, \text{SE}_i, D_i)$ does not depend on $D_i$. In this region, Equation 11 says that $D_i \in \mathcal{D}$. And then Equation 12 says that if $D_i \in \mathcal{D}$, then $\mu_i|r_i, \text{SE}_i$ does not depend on $D_i$. Thus, $E(\mu_i|r_i, \text{SE}_i, D_i)$ does not depend on $\text{SE}_i$ or $D_i$, and we can write, for $r_i > h$,

$$E(\mu_{\mathcal{D}}|r_i, \text{SE}_i, D_i) = E(\mu_{\mathcal{D}}|r_i) \tag{22}$$

where $\mu_{\mathcal{D}}$ is a r.v. generated by $g_{\text{SE}}(\cdot)$.

Now we just need to show that for $r_i > h$, $E(\mu_{\mathcal{D}}|r_i)$ is strictly increasing in $r_i$. Tweedie's Formula (Efron (2011) Equation (2.8)) implies

$$E\left(\mu_{\mathcal{D}}|r_i\right) = r_i + \text{SE}^2 \frac{d}{dr_i} \log f\left(r_i\right) \tag{23}$$

$$\text{Var}\left(\mu_{\mathcal{D}}|r_i\right) = \text{SE}^2 \left(1 + \text{SE}^2 \frac{d^2}{dr_i^2} \log f\left(r_i\right)\right) \tag{24}$$

where $f(r_i)$ is the marginal density of $r_i$. Differentiating the first equation with respect to $r_i$ and plugging in the second equation yields

$$\frac{d}{dr_i} E\left(\mu_{\mathcal{D}}|r_i\right) = \frac{\mathrm{Var}\left(\mu_{\mathcal{D}}|r_i\right)}{\mathrm{SE}^2} > 0 \tag{25}$$

where the inequality comes from the fact that $g_{\mathrm{SE}}(\cdot)$ has positive variance.

Thus, if $\mathrm{Var}(\mu_{\mathcal{D}}|r_i)$ is non-zero, then $E\left(\mu_{\mathcal{D}}|r_i\right)$ is a strictly increasing function of $r_i$. □

## B.2 Estimation Details

We construct the quasi-likelihood using the distr package in R (Ruckdeschel et al. (2006)) and optimize using the BOBYQA algorithm in the nloptr package (Johnson (2007)). BOBYQA is a derivative-free bound-constrained optimization based on quadratic approximations of the objective.

We also use distr to compute the prediction formula (Equation (4)). To ensure numerical stability, we split the integrals into many smaller parts.

## B.3 RW Method Details

This approach is closely related to $k$-family-wise error rate ($k$-FWER) control, as the FDP can be thought of as the number of family-wise errors divided by the number of discoveries.

We implement RW's Algorithm 4.1 as follows:

1. Let $h$ be the t-stat threshold from applying RW Algorithm 2.1 to control the $k$-FWER at level $q^*$.

2. Use bisection to find the largest $k$ such that

$$\frac{k}{[\text{Number of } |t_i| > h] + 1} > p^* \tag{26}$$

This method modifies RW's Algorithm 4.1 to be more computationally efficient. Instead of testing every $k = 1, 2, \ldots$ until Condition (26) is violated, we bisect to find this $k$ more quickly. As seen in the proof of Theorem 4.1 in RW, the Condition (26) ensures $\Pr(\mathrm{FDP} > p^*) \leq q^*$, while the sequence of $k$ examined does not matter.

We implement RW's Algorithm 2.1 as follows:

(a) Using all strategies, bootstrap 2,000 samples by demeaning returns at the strategy level and then re-sampling months with replacement. Let $|t_b^*|$ be the $k$th largest absolute t-stat across strategies in bootstrap $b$. Assign $h$ as the $(1-q^*)$ quantile of $|t_b^*|$ across the $B$ bootstraps.

(b) If $k=1$ then stop. If $\binom{\text{Number of strategies with } |t_i|>h}{k-1} > 100$ stop. Otherwise, repeat step (a) using the all possible sets of strategies that come from combining $k-1$ strategies with $|t_i|>h$ and all strategies with $|t_i| \leq h$. Update $h$ to be the largest hurdle across all possible sets.

(c) Repeat step (b) until $h$ does not change.

RW's Algorithm 2.1 does not have a stopping condition based on $\binom{\text{Number of strategies with } |t_i|>h}{k-1}$ but this condition is required to ensure that step (b) is computationally feasible. In fact, step (b) is typically infeasible in our setting, with tens of thousands of strategies. For example, if there are 1,000 strategies with $|t_i|>h$, and $k=3$, then there are $\binom{1000}{3} = 166$ million possible sets to consider in step (b). As seen in the proof of Theorem 2.1 in RW, imposing this stopping condition still ensures $k$-FWER control at level $q^*$.

# C Robustness

**Figure A.1: False Discovery Controls: Robustness** We repeat the exercise in Figure 6 using alternative parameter choices examined in Harvey, Liu, and Zhu (2016) (for BY 1.3), Barras, Scaillet, and Wermers (2010) (for Storey), and Harvey and Liu (2020) (for RW). **Interpretation:** As in Figure 6, the recommendations in Harvey et al. (2016) and Harvey et al. (2020) would lead one to miss most of the strategies with notable out-of-sample returns. Barras et al.'s recommendation performs much better.

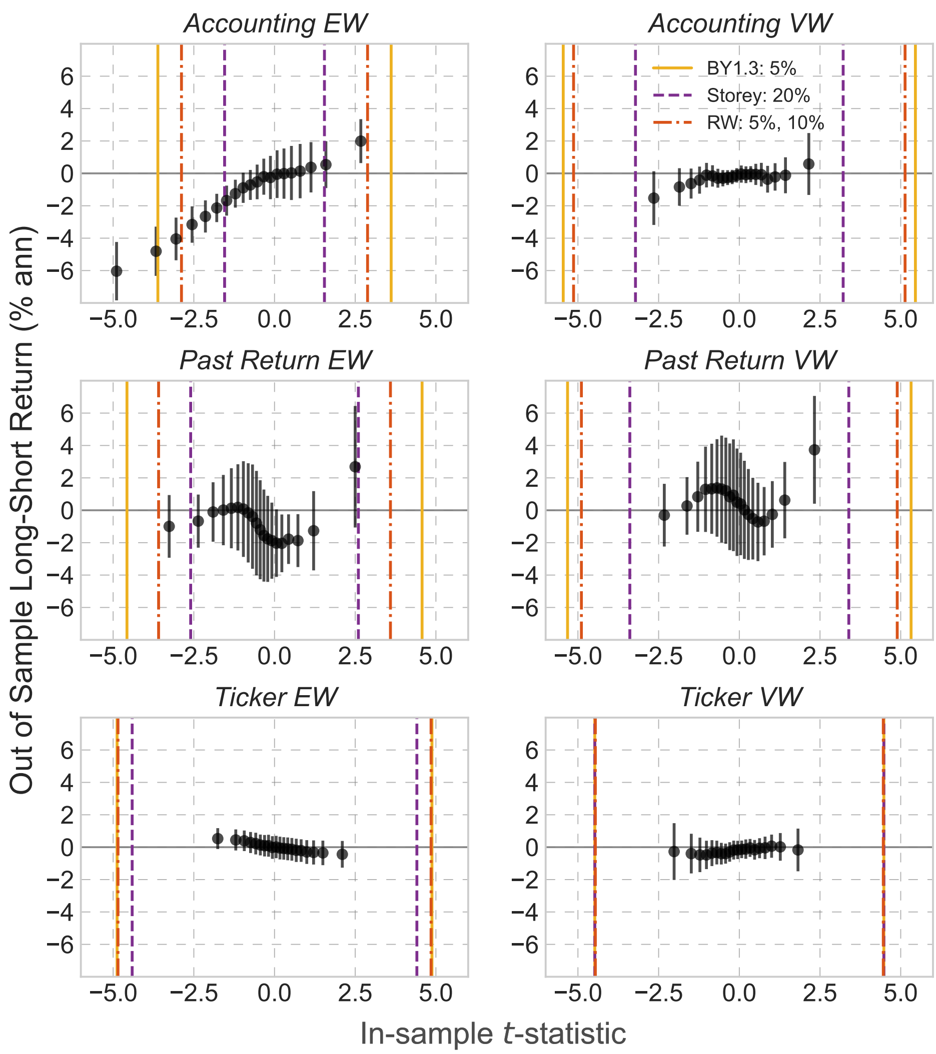


**Table A.1: Top 20 Strategies in 2003**

We repeat Table 2 Panel B using predicted Sharpe ratios from 1984-2003 and out-of-sample Sharpe ratios from 2004-2013. All strategies are equal-weighted. **Interpretation:** The list illustrates how the top strategies change over time. Shorting variables related to debt growth remains a common theme.

| Rank | Pred. SR (ann) | OOS SR (ann) | Signal Family | Signal Name |
|---|---|---|---|---|
| 1 | 1.60 | 0.64 | Acct | - LT Debt Issuance / Acc Depr, Depl & Amort |
| 2 | 1.53 | 1.27 | Acct | - ΔInterest & Rel Exp / Lag(Tot Liabilities) |
| 3 | 1.47 | 1.63 | Acct | - ΔInterest & Rel Exp / Lag(Tot Assets) |
| 4 | 1.46 | 0.90 | Acct | - ΔInterest & Rel Exp / Lag(Parent SH Equity) |
| 5 | 1.45 | 0.78 | Acct | - ΔNet Interest Paid / Lag(Parent SH Equity) |
| 6 | 1.45 | 0.87 | Acct | - ΔTot Liabilities / Lag(Tot Asset) |
| 7 | 1.43 | 1.38 | Acct | - ΔInterest & Rel Exp / Lag(Acc Depr, Depl & Amort) |
| 8 | 1.41 | 0.96 | Acct | - ΔTot Liabilities / Lag(Curr Liabilities) |
| 9 | 1.40 | 0.68 | Acct | - ΔTot Liabilities / Lag(Other Curr Assets) |
| 10 | 1.39 | 0.71 | Acct | - ΔNet Interest Paid / Lag(Comm Equity) |
| 11 | 1.39 | 0.81 | Acct | - ΔInterest & Rel Exp / Lag(Comm Equity) |
| 12 | 1.38 | 1.57 | Acct | - ΔInterest & Rel Exp / Lag(Invested Capital) |
| 13 | 1.38 | 1.15 | Acct | - Sale Comm & Pref Stk / Cash & ST Inv |
| 14 | 1.38 | 0.64 | Acct | - ΔTot Liabilities / Lag(Tot Liabilities) |
| 15 | 1.38 | 0.17 | Acct | - Mortg, Other Sec Debt / Acc Depr, Depl & Amort |
| 16 | 1.37 | 0.80 | Acct | - ΔInterest & Rel Exp / Lag(Comm Equity Liq Val) |
| 17 | 1.37 | 0.99 | Acct | - ΔLongterm Debt/ Lag(Acc Depr, Depl & Amort) |
| 18 | 1.37 | 0.75 | Acct | - ΔTot Liabilities / Lag(Tot Curr Assets) |
| 19 | 1.37 | 0.58 | Acct | - ΔTot Longterm Debt / Lag(Other Curr Assets) |
| 20 | 1.36 | 0.43 | Acct | - ΔTot Longterm Debt / Lag(Com Equity Tangible) |

**Table A.2: Top 20 Strategies in 2013**

We repeat Table 2 Panel B using predicted Sharpe ratios from 1994-2013 and out-of-sample Sharpe ratios from 2014-2020. All strategies are equal-weighted. **Interpretation:** Shorting variables related to debt growth remains a common theme.

| Rank | Pred. SR (ann) | OOS SR (ann) | Signal Family | Signal Name |
|---|---|---|---|---|
| 1 | 1.40 | 0.49 | Acct | - ΔInterest & Rel Exp / Lag(Total Assets) |
| 2 | 1.35 | 0.29 | Acct | - ΔInterest & Rel Exp / Lag(Invested Capital) |
| 3 | 1.25 | -0.01 | Acct | - ΔInterest & Rel Exp / Lag(Acc Depr, Depl & Amort) |
| 4 | 1.22 | 0.49 | Acct | - Financing Actv, Net Cash / Cash & ST Inv |
| 5 | 1.20 | 0.52 | Acct | - ΔInterest & Rel Exp / Lag(Market Val Equity) |
| 6 | 1.18 | 0.34 | Acct | - ΔInterest & Rel Exp / Lag(Tot Liabilities) |
| 7 | 1.18 | 0.10 | Acct | - ΔTot Liabilities / Lag(Total Assets) |
| 8 | 1.17 | 0.32 | Acct | - ΔNet PPE / Lag(Dep & Amort) |
| 9 | 1.16 | 0.62 | Acct | - ΔCost Goods Sold / Lag(Cost Goods Sold) |
| 10 | 1.16 | -0.37 | Acct | - ΔNet Interest Paid / Lag(Acc Depr, Depl & Amort) |
| 11 | 1.15 | 0.60 | Acct | - Sale Comm & Pref Stk / Cash & ST Inv |
| 12 | 1.14 | 0.17 | Acct | - ΔNet Interest Paid / Lag(Total Assets) |
| 13 | 1.14 | -0.37 | Acct | - ΔInterest & Rel Exp / Lag(Gross PPE) |
| 14 | 1.14 | -0.11 | Acct | - ΔInterest & Rel Exp / Lag(Dep & Amort) |
| 15 | 1.13 | -0.32 | Acct | - ΔTot Longterm Debt / Lag(Total Assets) |
| 16 | 1.13 | 0.26 | Acct | - ΔTot Liabilities / Lag(Curr Liabilities) |
| 17 | 1.13 | -0.20 | Acct | - ΔTot Longterm Debt / Lag(Acc Depr, Depl & Amort) |
| 18 | 1.13 | 0.19 | Acct | - ΔInterest & Rel Exp / Lag(Parent SH Equity) |
| 19 | 1.11 | 0.12 | Acct | - ΔInterest & Rel Exp / Lag(Capital Exp) |
| 20 | 1.11 | 0.42 | Acct | - Financing Actv, Net Cash / Market Val Equity |

## References


Barras, Laurent, Olivier Scaillet, and Russ Wermers (2010). “False discoveries in mutual fund performance: Measuring luck in estimated alphas”. In: *The journal of finance* 65.1, pp. 179–216.

Benjamini, Yoav (2010). “Discovering the false discovery rate”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 72.4, pp. 405–416.

— (2020). “Selective inference: The silent killer of replicability”. In.

Benjamini, Yoav and Yosef Hochberg (1995). “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal statistical society: series B (Methodological)* 57.1, pp. 289–300.

— (2000). “On the adaptive control of the false discovery rate in multiple testing with independent statistics”. In: *Journal of educational and Behavioral Statistics* 25.1, pp. 60–83.

Benjamini, Yoav, Abba M Krieger, and Daniel Yekutieli (2006). “Adaptive linear step-up procedures that control the false discovery rate”. In: *Biometrika* 93.3, pp. 491–507.

Benjamini, Yoav and Daniel Yekutieli (2001). “The control of the false discovery rate in multiple testing under dependency”. In: *Annals of statistics*, pp. 1165–1188.

— (2005). “False discovery rate–adjusted multiple confidence intervals for selected parameters”. In: *Journal of the American Statistical Association* 100.469, pp. 71–81.

Bodt, Eric de, B. Espen Eckbo, and Richard W. Roll (2025). “Competition Shocks, Rival Reactions, and Stock Return Comovement”. In: *Journal of Financial and Quantitative Analysis*. Published online 18 February 2025. DOI: 10.1017/S0022109024000486.

Bowles, Boone et al. (2023). “Anomaly time”. In: *Available at SSRN 3069026*.

Chen, Andrew Y (2024a). “Do t-Statistic Hurdles Need to be Raised?” In: *Management Science*.

— (2024b). “Most claimed statistical findings in cross-sectional return predictability are likely true”. In: *arXiv preprint arXiv:2206.15365*.

Chen, Andrew Y, Alejandro Lopez-Lira, and Tom Zimmermann (2022). “Peer-reviewed theory does not help predict the cross-section of stock returns”. In: *arXiv preprint arXiv:2212.10317*.

Chen, Andrew Y and Mihail Velikov (2022). "Zeroing in on the Expected Returns of Anomalies". In: *Journal of Financial and Quantitative Analysis*.

Chen, Andrew Y and Tom Zimmermann (2020). "Publication bias and the cross-section of stock returns". In: *The Review of Asset Pricing Studies* 10.2, pp. 249–289.

— (2022). "Open Source Cross Sectional Asset Pricing". In: *Critical Finance Review*.

Chordia, Tarun, Amit Goyal, and Alessio Saretto (2020). "Anomalies and false rejections". In: *The Review of Financial Studies* 33.5, pp. 2134–2179.

Chordia, Tarun, Avanidhar Subrahmanyam, and Qing Tong (2014). "Have capital market anomalies attenuated in the recent era of high liquidity and trading activity?" In: *Journal of Accounting and Economics* 58.1, pp. 41–58.

Cochrane, John H (2005). "The risk and return of venture capital". In: *Journal of financial economics* 75.1, pp. 3–52.

Didisheim, Antoine et al. (2023). *Complexity in factor pricing models*. Tech. rep. National Bureau of Economic Research.

Efron, B and T Hastie (2016). *Computer age statistical inference: Data mining, inference and prediction*. Cambridge: Cambridge University Press.

Efron, Bradley (2011). "Tweedie's formula and selection bias". In: *Journal of the American Statistical Association* 106.496, pp. 1602–1614.

— (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Vol. 1. Cambridge University Press.

Efron, Bradley and Carl Morris (1973). "Stein's estimation rule and its competitors-an empirical Bayes approach". In: *Journal of the American Statistical Association* 68.341, pp. 117–130.

Efron, Bradley, Robert Tibshirani, et al. (2001). "Empirical Bayes analysis of a microarray experiment". In: *Journal of the American statistical association* 96.456, pp. 1151–1160.

Engelberg, Joseph et al. (2023). "Do Cross-Sectional Predictors Contain Systematic Information?" In: *Journal of Financial and Quantitative Analysis* 58.3, pp. 1172–1201. DOI: 10.2139/ssrn.3459229.

Fama, Eugene F and Kenneth R French (1992). "The cross-section of expected stock returns". In: *the Journal of Finance* 47.2, pp. 427–465.

— (1993). "Common risk factors in the returns on stocks and bonds". In: *Journal of financial economics* 33.1, pp. 3–56.

Fisher, RA (1925). "Statistical methods for research workers." In.

Genovese, Christopher R, Kathryn Roeder, and Larry Wasserman (2006). "False discovery control with p-value weighting". In: *Biometrika* 93.3, pp. 509–524.

Harvey, Campbell R (2017). "Presidential address: The scientific outlook in financial economics". In: *The Journal of Finance* 72.4, pp. 1399–1440.

Harvey, Campbell R and Yan Liu (2020). "False (and missed) discoveries in financial economics". In: *The Journal of Finance* 75.5, pp. 2503–2553.

Harvey, Campbell R, Yan Liu, and Alessio Saretto (2020). "An evaluation of alternative multiple testing methods for finance applications". In: *The Review of Asset Pricing Studies* 10.2, pp. 199–248.

Harvey, Campbell R, Yan Liu, and Heqing Zhu (2016). "... and the cross-section of expected returns". In: *The Review of Financial Studies* 29.1, pp. 5–68.

Heath, Davidson et al. (2023). "Reusing Natural Experiments". In: *Journal of Finance* 78.4, pp. 2329–2364. DOI: `10.1111/jofi.13250`.

Jegadeesh, Narasimhan and Sheridan Titman (1993). "Returns to buying winners and selling losers: Implications for stock market efficiency". In: *The Journal of finance* 48.1, pp. 65–91.

Jensen, Theis Ingerslev, Bryan Kelly, and Lasse Heje Pedersen (2023). "Is there a replication crisis in finance?" In: *The Journal of Finance* 78.5, pp. 2465–2518.

Johnson, Steven G. (2007). *The NLopt nonlinear-optimization package*. `https://github.com/stevengj/nlopt`.

Kelly, Bryan, Semyon Malamud, and Kangying Zhou (2024). "The virtue of complexity in return prediction". In: *The Journal of Finance* 79.1, pp. 459–503.

Kim, Yong Hyuck, Zoran Ivkovich, and Dmitriy Muravyev (2021). "Causal Effect of Information Costs on Asset Pricing Anomalies". In: *Available at SSRN 3921785*.

Lakonishok, Josef, Andrei Shleifer, and Robert W Vishny (1994). "Contrarian investment, extrapolation, and risk". In: *The journal of finance* 49.5, pp. 1541–1578.

Marrow, Benjamin and Stefan Nagel (2024). *Real-Time Discovery and Tracking of Return-Based Anomalies*. Tech. rep. Working Paper.

McLean, R David and Jeffrey Pontiff (2016). "Does academic research destroy stock return predictability?" In: *The Journal of Finance* 71.1, pp. 5–32.

Peng, Lin (2005). "Learning with information capacity constraints". In: *Journal of Financial and Quantitative Analysis* 40.2, pp. 307–329.

Robbins, Herbert (1956). “An Empirical Bayes Approach to Statistics”. In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* 3.1.

Romano, Joseph P and Michael Wolf (2005). “Stepwise multiple testing as formalized data snooping”. In: *Econometrica* 73.4, pp. 1237–1282.

— (2007). “Control of Generalized Error Rates in Multiple Testing”. In: *The Annals of Statistics*, pp. 1378–1408.

Ruckdeschel, P. et al. (May 2006). “S4 Classes for Distributions”. English. In: *R News* 6.2, pp. 2–6.

Storey, John D (2002). “A direct approach to false discovery rates”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 64.3, pp. 479–498.

Storey, John D, Jonathan E Taylor, and David Siegmund (2004). “Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66.1, pp. 187–205.

Storey, John D and Robert Tibshirani (2001). *Estimating false discovery rates under dependence, with applications to DNA microarrays.* Tech. rep. Technical Report 2001-28, Department of Statistics, Stanford University.

Yan, Xuemin Sterling and Lingling Zheng (2017). “Fundamental analysis and the cross-section of stock returns: A data-mining approach”. In: *The Review of Financial Studies* 30.4, pp. 1382–1423.

Yang, Liangliang et al. (2021). “High-throughput methods in the discovery and study of biomaterials and materiobiology”. In: *Chemical reviews* 121.8, pp. 4561–4677.

# Figures

**Figure 1: Cumulative Long-Short Returns from Rigorous Data-Mining.** Each year, we sign strategies to have positive returns and form portfolios that equal-weight the top $X$% of predicted Sharpe ratios. "EB Mining" uses Equation (4) while "Naive Mining" uses the standard calculation. We hold for one year and repeat. **Interpretation:** Like published strategies, data-mined strategies show little cyclicality. Both published and data-mined strategies experience a break around the early-2000s, around the time when internet access became widespread.

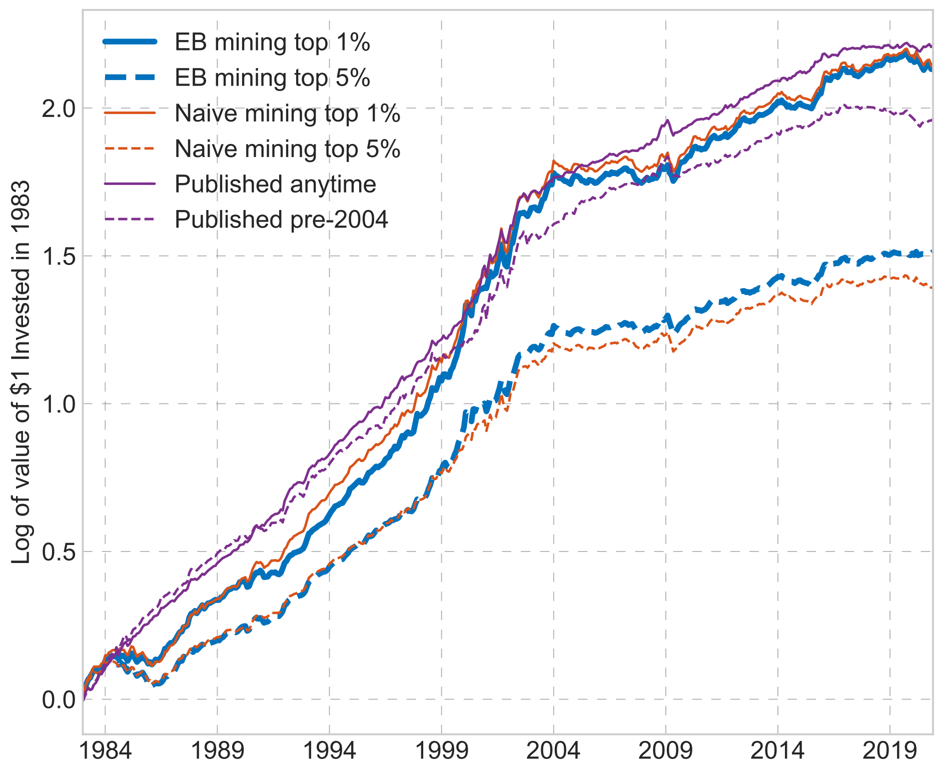


**Figure 2: Distribution of t-stats from long-short deciles strategies: 1983.** "Data" are t-stats testing the null of expected return = 0 from 1964-1983 for 136,000 trading strategies (Table 1). "Model" is Equations (13)-(14). "Null" is a standard normal. "EW" and "VW" are equal- and value-weighting, respectively. **Interpretation:** Equal-weighted accounting and equal-weighted past return strategies are far from the null, indicating true predictability. The models fit the data well.

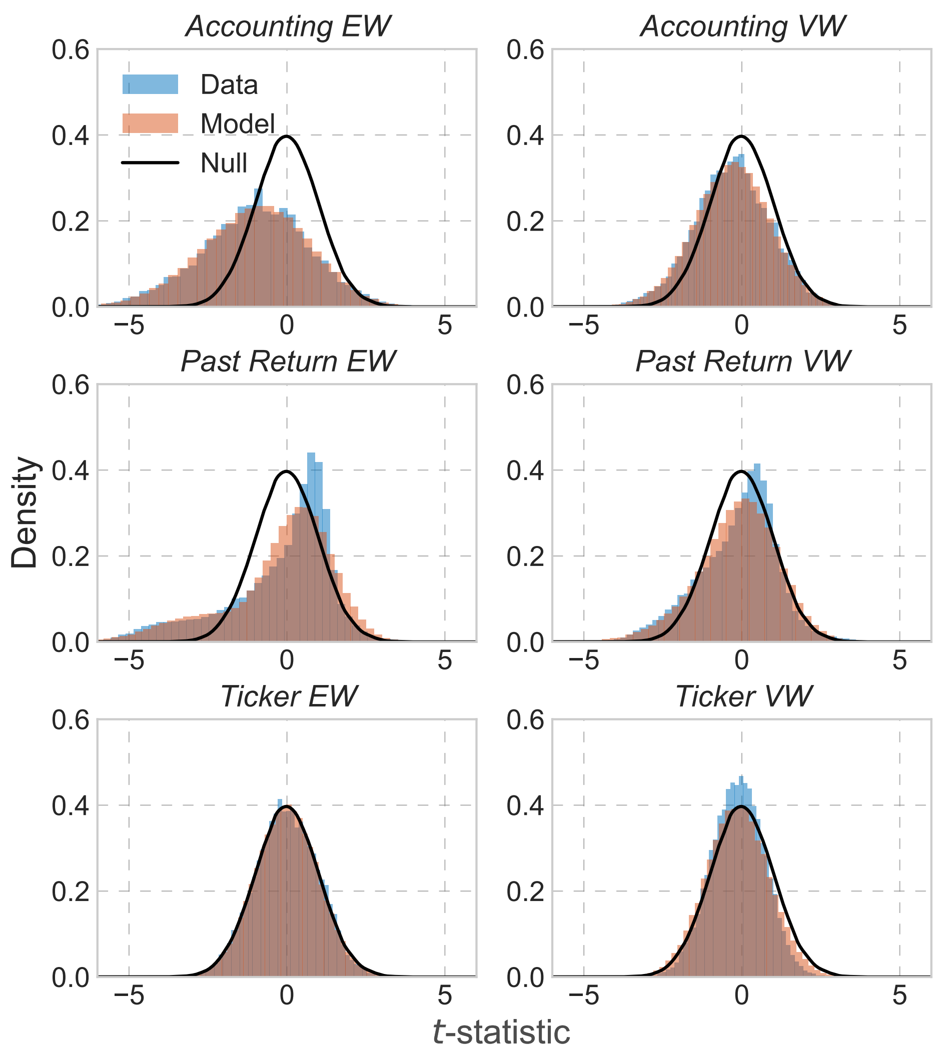


**Figure 3: Distribution of t-stats from long-short deciles strategies: 2004.** "Data" are t-stats testing the null of expected return = 0 from 1985-2004 for 136,000 trading strategies (Table 1). "Model" is Equations (13)-(14). "Null" is a standard normal. "EW" and "VW" are equal- and value-weighting, respectively. **Interpretation:** Compared to 1983 (Figure 2), t-stats from 2004 are much closer to the null, indicating diminished predictability. Equal-weighted accounting strategies still show true predictability, however.

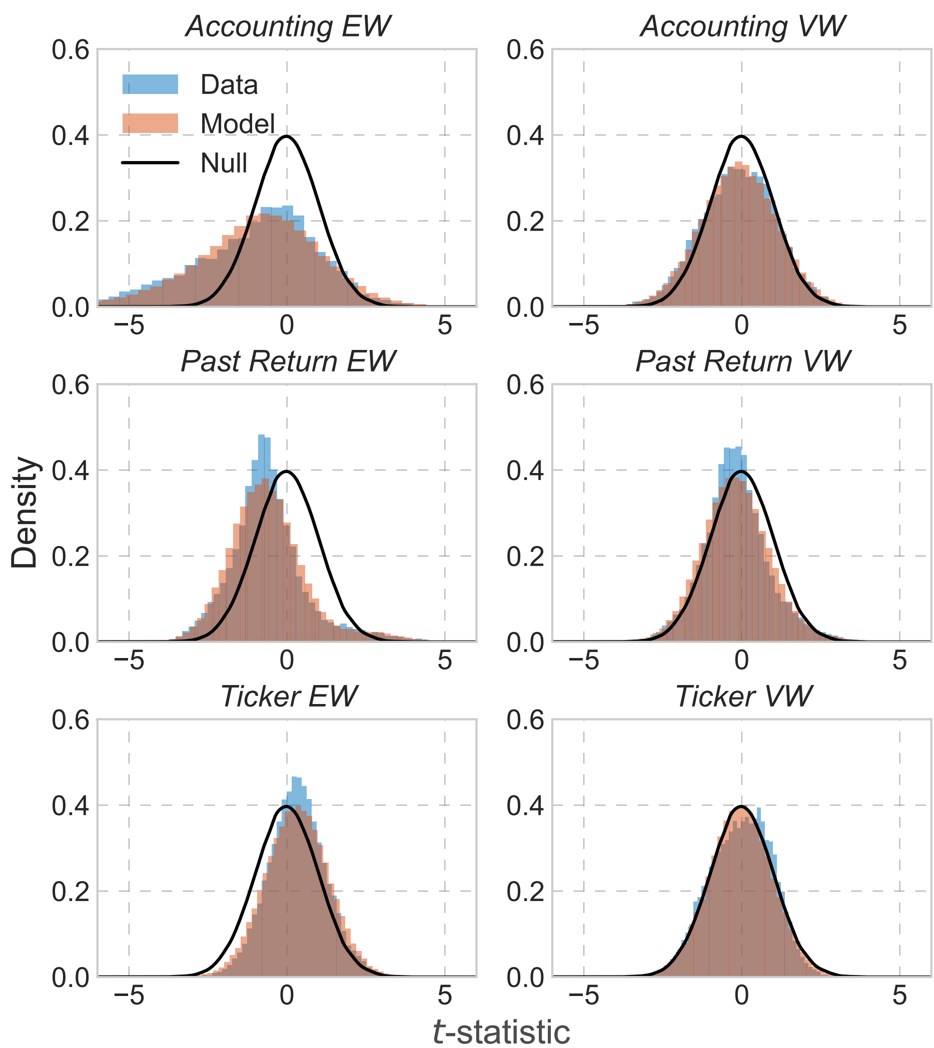


**Figure 4: Empirical Bayes Predictions and Out-of-Sample Returns: 1983-2004.** For each year and each family of strategies, we sort strategies into 20 groups based on the past 20 years of returns ("In-Samp") and predict returns using Bayes rule (Equation (3), "Predicted"). We form equal-weighted portfolios of strategies in each group and hold for one year ("OOS," error bars are two standard errors). **Interpretation:** Pre-2004, empirical Bayes shrinkage provides accurate forecasts of out-of-sample returns, unlike using the naive rule of in-sample return = out-of-sample return. Rigorous data mining removes data mining bias.

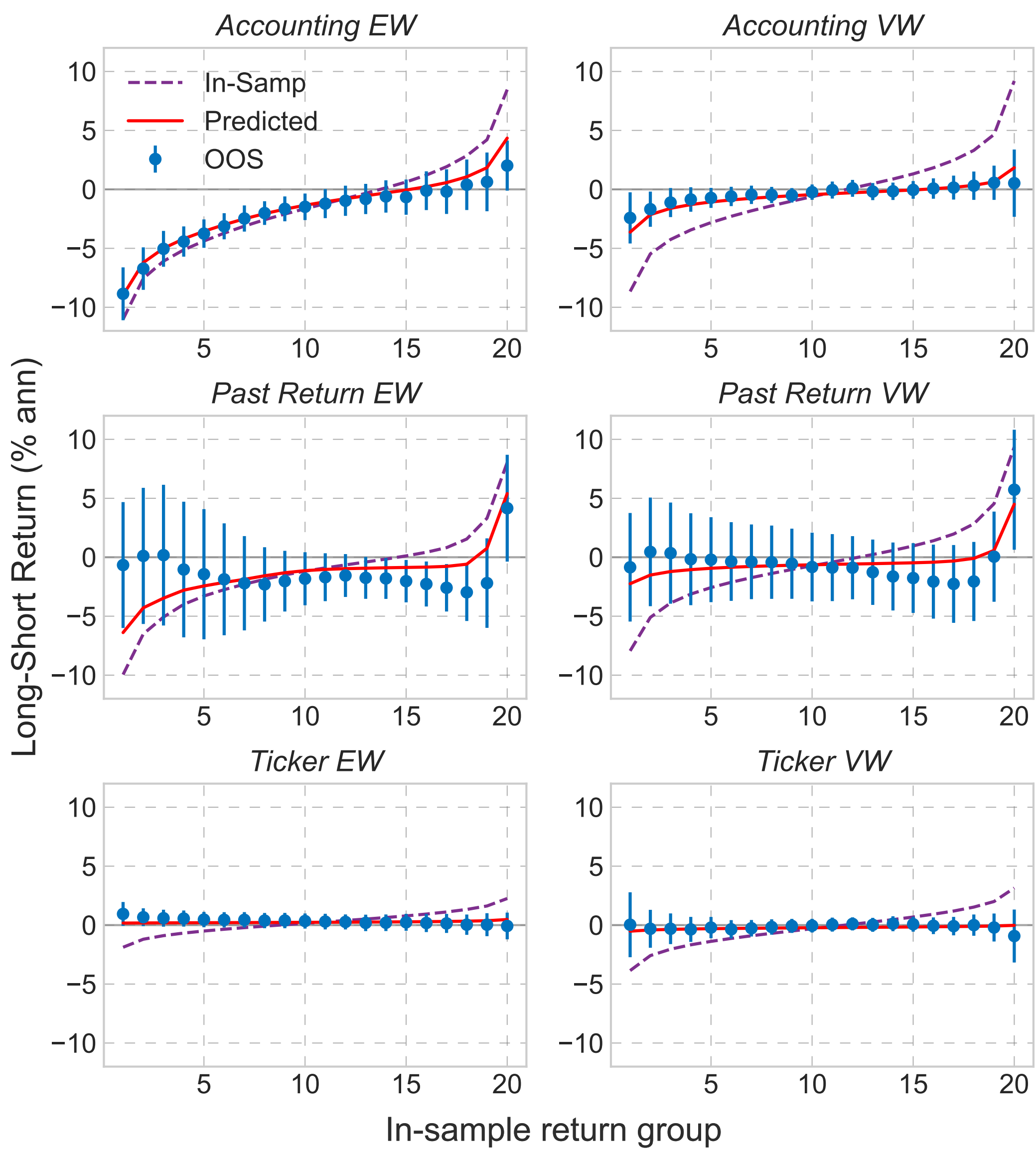


**Figure 5: Empirical Bayes Predictions and Out-of-Sample Returns: 2004-2020.** For each year and each family of strategies, we sort strategies into 20 groups based on the past 20 years of returns ("In-Samp") and predict returns using Bayes rule (Equation (3), "Predicted"). We form equal-weighted portfolios of strategies in each group and hold for one year ("OOS," error bars are two standard errors). **Interpretation:** Compared with pre-2004 (Figure 4), post-2004 predicted returns are closer to zero. Out-of-sample returns are even closer to zero, consistent with a structural break in predictability around 2004.

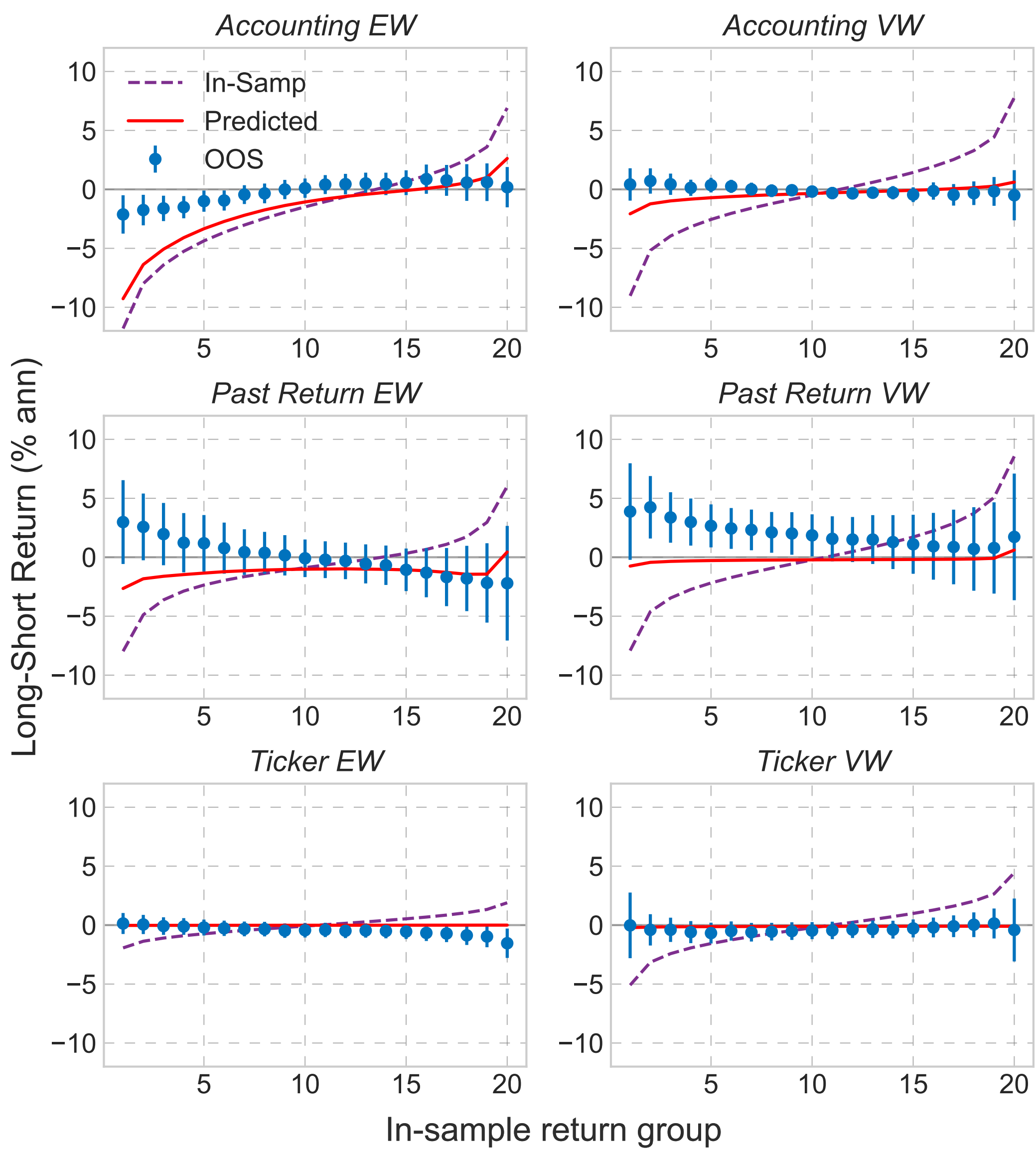


**Figure 6: False Discovery Controls** For each year and each strategy family, we calculate t-stat hurdles (vertical lines) using the recommendations of Harvey, Liu, and Zhu (2016) (BY 1.3: 1%); Barras, Scaillet, and Wermers (2010) (Storey: 10%); and Chordia, Goyal, and Saretto (2020) (RW: 5%, 5%). We compare with out-of-sample returns of strategies sorted into 20 bins based on in-sample t-statistics (markers). Hurdles, in-sample t-stats, and out-of-sample returns are calculated each year from 1983-2020, and then averaged across years. Error bars are two standard errors. **Interpretation:** Following the recommendations of Harvey et al. and Chordia et al. would lead one to miss most of the strategies with notable out-of-sample returns. The recommendation of Barras et al. performs much better.

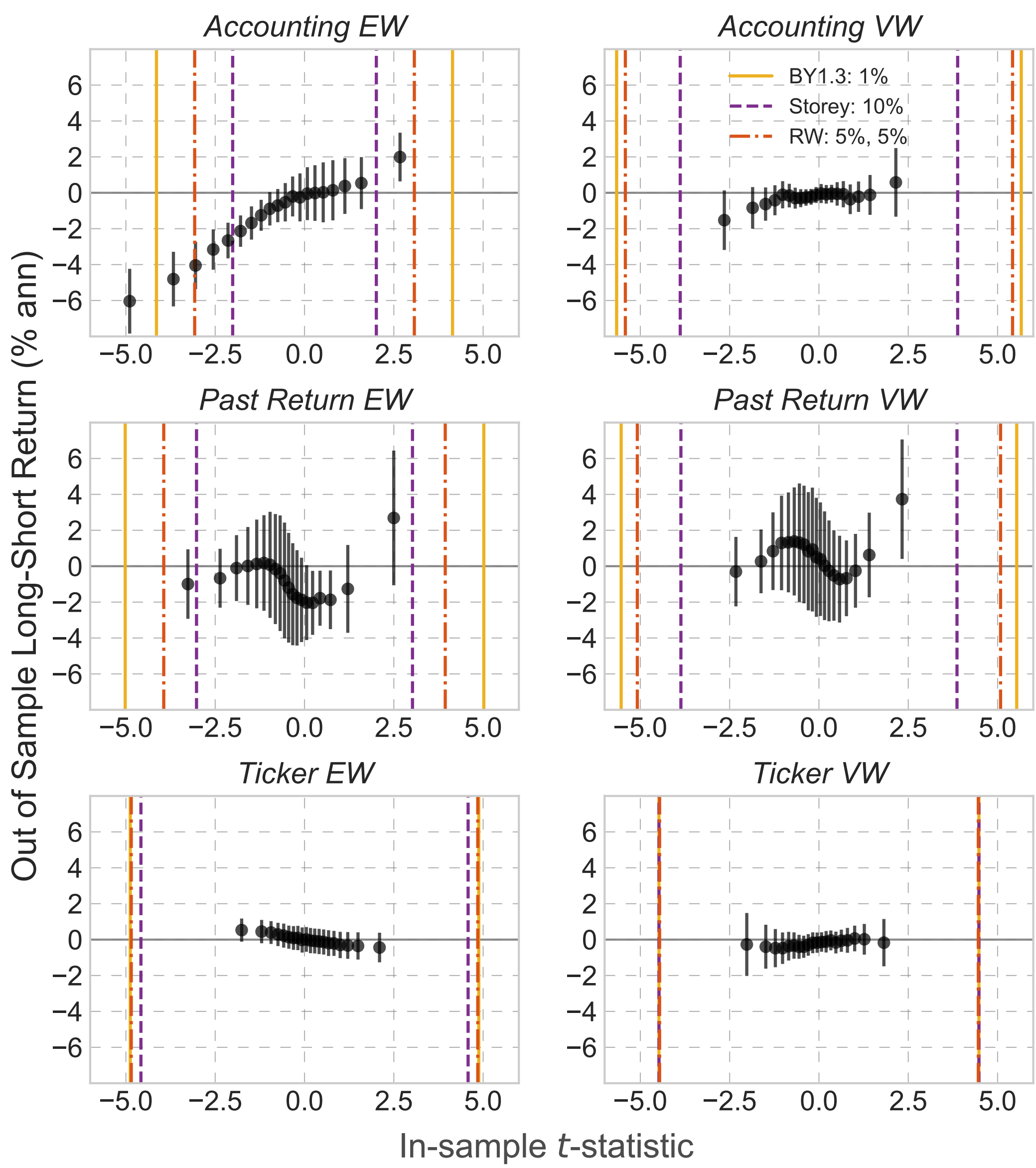
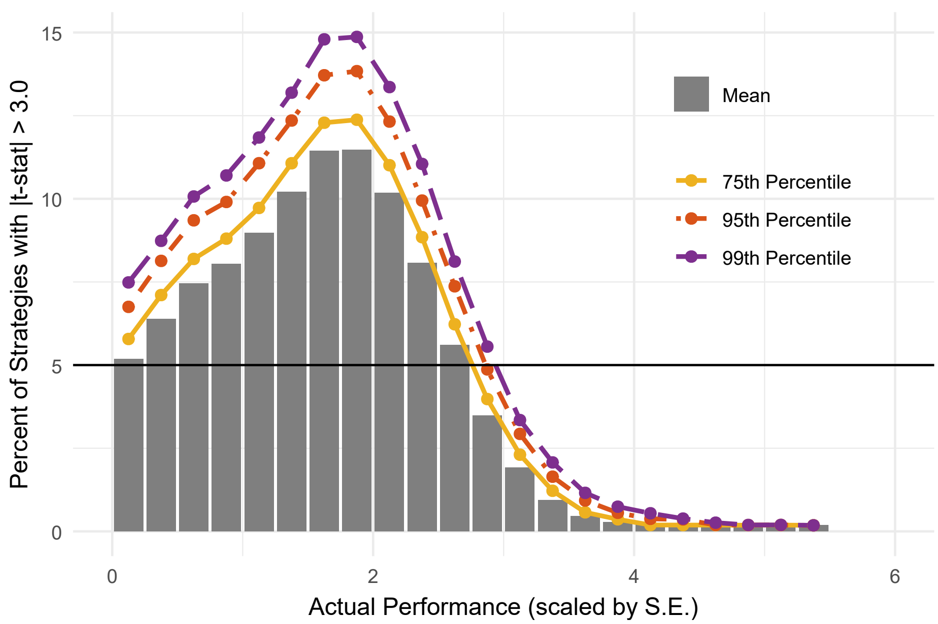


**Figure 7: FDP Risk Control Illustration.** Using the QML estimates for value-weighted accounting strategies based on data from 1964-1983, we run 2,000 simulations of 29,000 strategies, filter for $|t_i| > 3.0$, and then calculate histogram counts. For simplicity, we assume all signals are independent. The plot shows various statistics for each histogram bin, calculated across simulations. The FDP is approximately the share of strategies in the first bin. **Interpretation:** On average, the FDP is 5%, meaning the FDR is approximately controlled with a hurdle of 3.0. However, there is a risk that FDP > 5%, and thus a hurdle > 3.0 is needed to ensure Pr(FDP > 5%) < 5%. Since the FDP itself measures a left tail, the RW method aims to control the tail risk of a tail risk. This consevatism only makes sense if selecting a null strategy is catastrophic.



# Tables

**Table 1: Overview of 136,000 Long-Short Strategies**

Table describes the 136,192 strategies used throughout the paper. Data and code for these strategies are posted publicly. **Interpretation:** Unlike datasets of published strategies (e.g. Chen and Zimmermann (2022)), these strategies are arguably constructed without data-mining bias.

Panel A: Accounting Strategies

Description: Make ratios from 242 accounting variables by (1) dividing one variable by another and (2) taking first differences and then dividing. Long / short the extreme deciles. Data is from Chen, Lopez-Lira, and Zimmermann (2022).

| | # strategies | Mean Return (% ann) | | |
|---|---|---|---|---|
| | | 5 pctile | 50 pctile | 95 pctile |
| EW | 29,314 | -7.0 | -1.1 | 3.7 |
| VW | 29,314 | -4.5 | -0.4 | 3.9 |

Panel B: Past Return Strategies

Description: Choose 4 quarters out of the past 20 and compute one of the first four central moments, yielding $\binom{20}{4} \times 4 = 19,380$ signals. Add the return over any of the past 20 quarters and mean returns over the past 2 and past 3 quarters to arrive at 19,402 signals. Long / short the extreme deciles.

| | # strategies | Mean Return (% ann) | | |
|---|---|---|---|---|
| | | 5 pctile | 50 pctile | 95 pctile |
| EW | 19,402 | -5.3 | -0.4 | 2.1 |
| VW | 19,402 | -3.4 | 0.1 | 4.3 |

Panel C: Ticker Strategies

Description: Sort stocks into 20 groups based on alphabetical order of the first ticker symbol. Long two of those groups and short two. Repeat using the 2nd, 3rd, and 4th ticker symbols. This yields $\binom{20}{4} \times 4 = 19,380$ long-short portfolios.

| | # strategies | Mean Return (% ann) | | |
|---|---|---|---|---|
| | | 5 pctile | 50 pctile | 95 pctile |
| EW | 19,380 | -0.9 | 0.0 | 0.8 |
| VW | 19,380 | -2.2 | -0.2 | 1.6 |

**Table 2: Returns of Data-Mined Long-Short Portfolios**

Each year, we sign strategies to have positive predicted returns and form portfolios that equally-weights the top $X$% of strategies based on their Sharpe ratios. “EB Mining” uses Equation (4) while “Naive Mining” uses the standard calculation. We hold for one year and repeat. ’Pub Anytime’ is a portfolio that equally weights strategies from Chen and Zimmermann (2022). ’Pub Pre-2004’ equally weighs strategies published before 2004. **Interpretation:** Rigorous data mining generates out-of-sample returns comparable to those from the best journals in finance, even if the data includes signals with zero out-of-sample mean returns, like ticker-sorted portfolios.

| | Num Strats Combined | Mean Return (% ann) | $t$-stat | Sharpe Ratio (ann) |
|---|---|---|---|---|
| Panel A | | | | |
| EB Mining Top 1% | 1278 | 5.70 | 9.00 | 1.46 |
| EB Mining Top 5% | 6389 | 4.03 | 8.27 | 1.34 |
| EB Mining Top 10% | 12777 | 2.77 | 7.16 | 1.16 |
| Panel B | | | | |
| Naive Mining Top 1% | 1278 | 5.72 | 8.91 | 1.45 |
| Naive Mining Top 5% | 6389 | 3.72 | 7.73 | 1.25 |
| Naive Mining Top 10% | 12777 | 2.61 | 6.77 | 1.10 |
| Panel C | | | | |
| Pub Anytime | 203 | 5.88 | 12.54 | 2.03 |
| Pub Pre-2004 | 82 | 5.23 | 9.57 | 1.55 |

**Table 3: Description of the Top 1% Data-Mined Strategies**

Panel A shows the fraction of strategies that comes from each signal family, pooled across all sample years. Panel B lists the definitions of the strategies with highest predicted Sharpe Ratios (SR pred) using data from 1974-1993. SR OOS is the realized Sharpe ratio 1994-2003. All strategies in Panel B are equal-weighted. **Interpretation:** The top 1% strategies are largely equal-weighted accounting strategies. Equal-weighted past return strategies comprise a non-trivial minority. The top 20 strategies are distant from strategies popular the academic literature at the time of Fama and French (1993), yet they perform well out-of-sample.

| Panel A: Average Fraction of Signals in the Top 1% | | | | | |
|---|---|---|---|---|---|
| Acct EW | Acct VW | Past Ret EW | Past Ret VW | Ticker EW | Ticker VW |
| 91.0% | 0.3% | 8.6% | 0.1% | 0.0% | 0.0% |

| Panel B: Top 20 Strategies in 1993 based on Signed Predicted Sharpe Ratio | | | | |
|---|---|---|---|---|
| Rank | SR Pred | SR OOS | Signal Family | Signal Name |
| 1 | 1.56 | 1.32 | Acct EW | - ΔInterest paid net / Lag(Common equity) |
| 2 | 1.51 | 0.84 | Acct EW | - Debt due in 2nd year / Depr, depl & amort |
| 3 | 1.43 | 0.92 | Acct EW | - Debt mortgages & other sec / Sales |
| 4 | 1.37 | 1.60 | Acct EW | - Debt mortgages & other sec / Depr, depl & amort |
| 5 | 1.37 | 1.64 | Acct EW | - ΔInterest paid net / Lag(Stockholders equity) |
| 6 | 1.35 | 0.54 | Past Ret EW | + Return in quarters $t$ minus 5, 9, 17, and 18 |
| 7 | 1.35 | 0.68 | Acct EW | - Debt due in 3rd year / Depr, depl, and amort |
| 8 | 1.35 | 0.69 | Acct EW | - Debt mortgages & other sec / Cost of goods sold |
| 9 | 1.34 | 1.00 | Acct EW | - ΔInterest paid net / Lag(Inventories) |
| 10 | 1.33 | 0.62 | Acct EW | - Debt mortgages & other sec / Operating expenses |
| 11 | 1.33 | 0.47 | Past Ret EW | + Return in quarters $t$ minus 17 |
| 12 | 1.32 | 0.62 | Past Ret EW | + Return in quarters $t$ minus 9, 17, 18 and 19 |
| 13 | 1.30 | 0.43 | Acct EW | - ΔLiabilities / Lag(Depr & amort) |
| 14 | 1.29 | 0.44 | Past Ret EW | + Return in quarters $t$ minus 9, 13, 17, and 18 |
| 15 | 1.29 | 1.24 | Acct EW | - ΔInterest paid net / Lag(Equity liquidation value) |
| 16 | 1.29 | 0.68 | Past Ret EW | + Return in quarters $t$ minus 3, 9, 17, and 18 |
| 17 | 1.25 | 0.49 | Acct EW | - Debt due in 4th year / Depr, depl & amort |
| 18 | 1.25 | 1.01 | Acct EW | - Stock issuance / Gross profit |
| 19 | 1.25 | 0.55 | Acct EW | - Debt due in 2nd year / Depr & amort |
| 20 | 1.24 | 1.51 | Acct EW | - ΔLiabilities / Lag(Depr, depl & amort) |