CFR Working Paper No. 24-02

Does Peer-Reviewed Research Help
Predict Stock Returns?

A. Y. Chen ● A. Lopez-Lira
● T. Zimmermann

centre for financial research
cologne

Does Peer-Reviewed Research Help Predict Stock Returns?

Andrew Y. Chen[1], Alejandro Lopez-Lira[2], and Tom Zimmermann[3, 4]

[1]Federal Reserve Board
[2]University of Florida
[3]University of Cologne
[4]Centre for Financial Research

April 2024

Abstract

Mining 29,000 accounting ratios for t-statistics over 2.0 leads to cross-sectional predictability similar to the peer review process. For both methods, about 50% of predictability remains after the original sample periods. Data mining generates other features of peer review including the rise in returns as original sample periods end, the speed of post-sample decay, and themes like investment, issuance, and accruals. Predictors supported by peer-reviewed risk explanations underperform data mining. Similarly, the relationship between modeling rigor and post-sample returns is negative. Our results suggest peer review systematically mislabels mispricing as risk, though only 18% of predictors are attributed to risk.

1 Introduction

Suppose a Ph.D. student tells you he found a predictor with a long-short return of 100 bps per month in a historical sample. You ask him, "where does this predictor come from?" How would your view about the post-sample return change if the predictor is:

1. Based on an idea that is publishable in a top finance journal (e.g. Journal of Finance)

2. Found by mining tens of thousands of accounting ratios for t-stats greater than 2.0?

One would think the publishable predictor has a higher post-sample return. To be publishable, a predictor needs much more support than a t-stat greater than 2.0: it needs robustness tests, supporting evidence, and theoretical justification, at least most of the time. Some might expect the publishable predictor to have a *much* higher return, given the pressures and rewards of academic finance (Celerier, Vallee, and Vasilenko (2022)), and the warnings about data mining found throughout the asset pricing literature (Jensen and Benington (1970); Sullivan, Timmermann, and White (2001)). But how much higher is the publishable post-sample return? In other words, how much does peer-reviewed research help predict cross-sectional returns compared to data mining?

To answer this question, we construct the empirical counterpart to the scenario. We match 200 published predictors (from Chen and Zimmermann (2022)) to data-mined benchmarks. The data-mined benchmarks come from searching 29,000 accounting ratios for t-stats greater than 2.0 in the published predictors' original sample periods. The accounting ratios are naive: they are simply ratios or scaled first differences using 240 Compustat accounting variables (+ CRSP market equity). The only restriction on these ratios is that we avoid dividing by variables that are typically zero or negative. We form long-short portfolios for each predictor and re-scale so that the mean original-sample return is 100 bps per month. Finally, we compare post-sample returns.

Figure 1 illustrates the result. It plots the trailing 5-year return in event time, where the event is the month that the original sample ended. As shown in the seminal McLean and Pontiff (2016) meta-study, published returns (solid line) decay post-sample, but they remain far above zero, averaging 53% of their original sample means. Data-mined returns (dashed line) decay a bit more, with post-sample means that are 51% of their original sample means. So peer-reviewed research seems to help predict returns compared to data mining but the improvement is modest. The publishable predictor in our hypothetical scenario outperforms by only 2 bps per month.

In fact, Figure 1 shows that it is hard to reject the null that the academic predictor discovery process is, itself, data mining. Data mined returns match not only the post-

Figure 1: Does Peer-Reviewed Research Help Predict Returns?
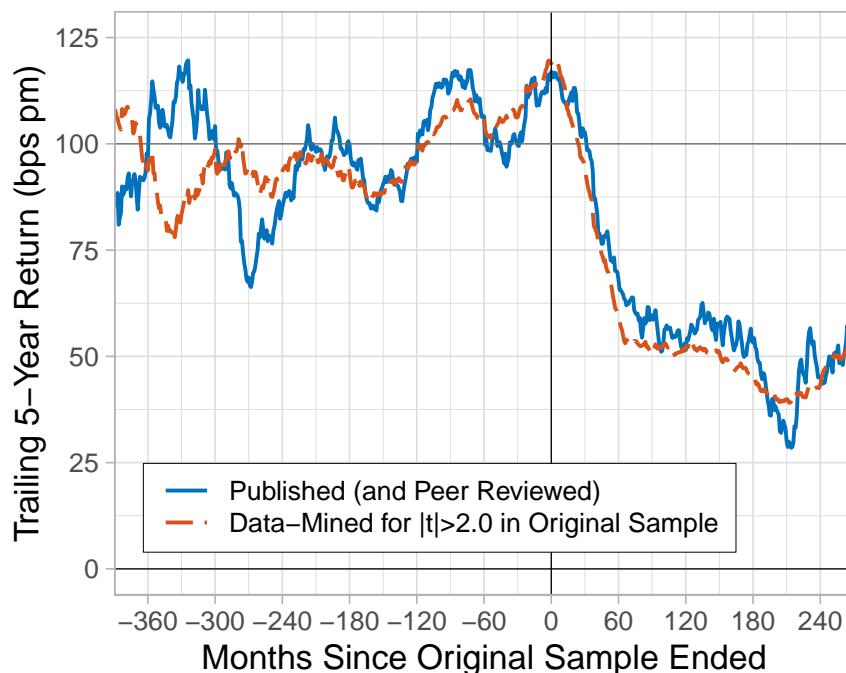
sample decay of published returns: they also match the rise in trailing 5-year returns as the original samples end, the decline in returns in the first 60 months post-sample, the flattening of returns in months 60 to 120, and even the dip in returns around month 210.

Perhaps the published predictors would outperform if we focused on papers that use risk-based ideas. As described in Cochrane's (2009) influential textbook, "the best hope for finding pricing factors that are robust out of sample and across different markets, is to try to understand the fundamental macroeconomic sources of risk." Many of the papers in the Chen and Zimmermann (2022) dataset do not follow this advice, and motivate their predictors using informal arguments about mispricing. Some even lack a clear explanation, and base their conclusions on the strength of their empirical results. For example, Banz (1981) ends with "the size effect exists but it is not at all clear why it exists."

To address this possibility, we assign predictors to "risk," "mispricing," or "agnostic" groups based on the explanation for predictability in the original papers. We then compare the post-sample returns of each group.

The main result remains: risk-based research does not lead to higher post-sample returns compared to data mining. If anything, risk-based predictors underperform their data mined benchmarks. We find similar results when categorizing predictors based on the support of a mathematical equilibrium model. While there are relatively few

predictors supported by formal models, the ones that exist imply that the relationship between modeling rigor and post-sample returns is *negative*.

An important caveat is that our results characterize cross-sectional predictability research as it was conducted from 1980 to 2016, the publication years covered by the Chen and Zimmermann (2022) dataset. Research evolves over time, and since 2016, a growing number of academics have embraced machine learning and other big data methods (Moritz and Zimmermann (2016); Yan and Zheng (2017); Gu, Kelly, and Xiu (2020)).[1] Indeed, fields like protein folding and language modeling have been recently revolutionized by atheoretical searches through vast amounts of data (Jumper et al. (2021); Zhao et al. (2023)).

Figure 1 is a stark illustration of the promise of big data methods. Simply searching accounting variables for large t-statistics generates substantial out-of-sample returns. And while data-mined results say little about the underlying economics, they can provide the empirical foundation for the next generation of economic ideas (Chen and Dim (2023)).

As a secondary result, we document a striking consensus about the origins of cross-sectional predictability, according to peer review. Among the 199 published predictors we examine, only 18% are attributed to risk by peer review. 59% are attributed to mispricing, and 23% have uncertain origins.

This consensus is a positive sign regarding the scientific process in finance. The fact that risk-based predictors consistently decay post-sample implies that peer review either mislabels mispricing as risk or identifies unstable risk factors that weaken over time. Fortunately, these errors are uncommon, and represent a relatively small "false discovery rate."

A more negative view of peer review comes from the fact that recent reviews of cross-sectional predictability are agnostic about risk vs mispricing (Bali, Engle, and Murray (2016); Zaffaroni and Zhou (2022)). Given the strong consensus found from reading the individual papers, this agnosticism suggests that the battle between risk-based and behavioral finance has led to an unwillingness to engage in debate. This unwillingness raises questions about whether the field of asset pricing is self-correcting (Ankel-Peters, Fiala, and Neubauer (2023)).

The remainder of this section reviews literature. Section 2 compares research with data mining. Section 3 examines whether risk-based ideas improve outperformance.

---

[1]An incomplete list of additional big data papers includes Green, Hand, and Zhang (2017); DeMiguel et al. (2020); Freyberger, Neuhierl, and Weber (2020); Kozak, Nagel, and Santosh (2020); Han et al. (2022); Chen and Velikov (2022); Bessembinder, Burt, and Hrdlicka (2021); Lopez-Lira and Roussanov (2020); Jensen, Kelly, et al. (2022); and Chen and McCoy (2024).

Section 4 provides data-mined alternatives to B/M and momentum. Section 5 provides robustness tests. Section 6 concludes.

1.1 Related Literature

To our knowledge, our paper is the first to test the widely-held belief that economic theory improves out of sample robustness relative to data mining. In previous research, this belief is either assumed to be true (Harvey, Liu, and Zhu (2016); Harvey (2017); Fama and French (2018)) or expressed as a "best hope" (Cochrane (2009)). Our tests provide evidence inconsistent with this belief—if theory is practiced the way it was in the papers covered by the Chen and Zimmermann (2022) dataset. We also provide a meta-theory for understanding why theory may fail to improve robustness.

Earlier papers on data mining in asset pricing studied statistical theory (Lo and MacKinlay (1990), see also Chen (2021)) or data mining for aggregate predictability (Sullivan, Timmermann, and White (1999); Sullivan, Timmermann, and White (2001)). Our paper fits in with the more recent literature following on Yan and Zheng (2017), which examines data mining for cross-sectional predictability (Chordia, Goyal, and Saretto (2020); Harvey and Liu (2020); Goto and Yamada (2022); Zhu (2023); Chen (2024)). Relative to these papers, ours is unique in showing that data mining works about as well as peer-reviewed research. We are also unique in focusing on out-of-sample tests, which are well-understood and have straightforward interpretations. The aforementioned papers focus on multiple testing methods, which can be easily misinterpreted (Chen and Zimmermann (2023)). Following up on our paper, Chen and Dim (2023) show how to use empirical Bayes to mine more rigorously.

Quantifying peer-reviewed texts provides a new angle on the long-standing debate on risk vs mispricing in the cross-section of stock returns (Fama (1970); Shiller (2003); Cochrane (2017); Barberis (2018); etc). Since Fama (1970), it has been recognized that standard empirical tests can only reject special cases of the broad class of risk theories (the "joint hypothesis problem"). Our methods attack this problem by building on the efforts of the asset pricing community. This community is, in a way, an organic computer designed to search the entire class of risk theories. Based on our tests, this search has uncovered little robust cross-sectional risk during the years covered by the Chen and Zimmermann (2022) dataset. Our use of peer-reviewed judgments is also distinct from contemporaneous papers that aim to classify anomalies in terms of risk vs mispricing, which focus on option market volume (Böll et al. (2024)), analyst forecasts revisions (Frey (2023)), stochastic dominance (Holcblat, Lioui, and Weber (2022)), or optimized factor

models (Bali, Beckmeyer, and Wiedemann (2023)).

2 Research vs Data Mining

We describe how we mine accounting data (Section 2.1.1), show that data mining produces out-of-sample returns (Section 2.2), and compare with published predictors (Section 2.3). We also illustrate the importance of mining accounting data, rather than other kinds of data (Section 2.4).

2.1 Data

Our primary data consists of predictors, either mined from Compustat or replicated based on published papers. We measure the magnitude of predictability using long-short strategy returns.

2.1.1 Data-Mined Accounting Predictors

We generate 29,315 firm-level signals as follows. Let X be one of 242 Compustat accounting variables + CRSP market equity and Y be one of the 65 variables that is observed and positive for $> 25\%$ of firms in 1963 with matched CRSP data. The 242 variables are the ones examined by Yan and Zheng (2017), who select these variables to ensure non-missing values in at least 20 years and that the average number of firms with non-missing values is at least 1,000 per year. We form signals by combining all combinations of ratios (X/Y) and scaled first differences ($\Delta X/\text{lag}(Y)$). Restricting Y to be positive for at least a meaningful minority of stocks avoids normalizing by zero or negative numbers. This procedure would lead to $242 \times 65 \times 2 = 31,460$ signals, but we drop 2,145 signals that are redundant in "unsigned" portfolio sorts.[2]

We lag each signal by six months relative to the fiscal year ends, and then form long-short decile strategies by sorting stocks on the lagged signals in each June. Delisting returns and other data handling methods follow Chen and Zimmermann (2022). For further details, please see the Github repo.

In our view, this process is the simplest reasonable data mining procedure. A reasonable data mining procedure should include both ratios and first differences. Scaling first

---

[2]For the $65 \times 65 = 4,225$ ratios where the numerator is also a valid denominator, there are only 65 choose 2 = 2,080 ratios that are distinct in the sense that there are no ratios which would lead to identical rankings if the sign was flipped.

differences by a lagged variable nests percentage changes, which likely should also be included in a reasonable data mining process. This data mining procedure includes little, if any, economic insight.

This procedure is inspired by Yan and Zheng (2017), who create 18,000 signals by applying 76 transformations to 240 accounting variables. These transformations are inspired, in part, by the asset pricing literature. Choosing transformations based on the literature could potentially lead to look-ahead bias, which our procedure avoids. However, previous versions of this paper used Yan and Zheng's data and found very similar results.

2.1.2 Published Predictors

Peer-reviewed predictors come from the August 2023 release of the Chen and Zimmermann (2022) (CZ) dataset. This dataset is built from 212 firm-level variables that were shown to predict returns cross-sectionally. It covers the vast majority of firm-level predictors that can be created from widely-available data and were published before 2016.

We drop five predictors that produce mean long-short original-sample returns of less than 15 bps per month in CZ's replications. These predictors are rather distant from the original papers, and dropping them ensures that the decay we document accurately reflects the literature.[3] Since these predictors are rare, including them has little effect on our results.

We drop another 8 predictors that have less than 9 years of post-sample returns. Most of these predictors rely on specialized data that have been discontinued, though a few are published relatively recently. This filter makes the post-sample results easy to interpret. But since the median post-sample length is about 20 years, including these predictors has little effect on our results.

For measuring the magnitude of predictability, we use the "original paper" version of the CZ data. These data consist of long-short portfolios constructed following the procedures in the original papers. This choice is important, as post-sample decay varies by the details of the trading strategy (Chen and Velikov (2022)). Choosing the original implementations means that the decay we find is not due to a dispute with the peer review process about where exactly expected returns should show up.

---

[3]For example, CZ equal-weight the Frazzini and Pedersen (2014) betting against beta portfolios instead weighting by betas. CZ use CRSP age rather than the NYSE archive data used by Barry and Brown (1984). CZ also find very small returns in simple long-short strategies for select variables shown by Haugen and Baker (1996), Abarbanell and Bushee (1998), Soliman (2008) to predict returns in multivariate settings.

2.2 "Out-of-Sample" Returns from Data Mining

Our simple data mining procedure generates notable "out-of-sample" returns, as seen in Table 1. Each June, we sort the 29,000 data-mined strategies into five bins based on their mean returns over the past 30 years ("in-sample") and compute the mean return over the next year within each bin ("out-of-sample"). We then average these statistics across each year. We put "out-of-sample" in quotes because this concept differs from the post-sample concept used in the rest of the paper.

The equal-weighted bin 1 returns -59 bps per month in-sample, with an average t-stat of -4. These statistics are similar to the typical published predictor (Chen and Zimmermann (2022)). Out-of-sample, this bin returns -49 bps per month, implying a mild decay of only 17%. Since investors can flip the long and short legs of these strategies, these statistics imply substantial out-of-sample returns. Similar predictability is seen in bin 5, which decays by 27%. Bins 2 and 3 also show out-of-sample predictability, though the decay is larger.

Out-of-sample predictability is also seen in value-weighted strategies, though the magnitudes are weaker. Still, the roughly 60% decay is far from 100%, and is in the ballpark of the post-sample decay for published predictors (McLean and Pontiff (2016)).

Post-2004 (Panel (b)), predictability is much weaker. The equal-weighted bins 1 and 5 returns decay by 55% and 33%, respectively in the first year out-of-sample. Decay is close to 100% in the extreme value-weighted strategies. This decline in predictability is consistent with the idea that the rise of information technology in the early 2000's has significantly reduced mispricing. Panel (b) suggests that the post-2004 decay is found not only in published predictors (Chordia, Subrahmanyam, and Tong (2014); Chen and Velikov (2022); Bowles et al. (2023)), but throughout the equities market.

Since each bin consists of approximately 6,000 strategies, Table 1 implies that *thousands* of strategies have notable out-of-sample predictability. But are these strategies distinct? To address this question, we describe the covariance structure of data-mined strategies in Table 1. The table examines strategies that have t-stats greater than 2.0 in at least 10% of the 30-year in-sample periods from Table 1.

Table 2 shows that data-mined strategies are generally distinct. More than 80% of pairwise correlations are below 0.29 in absolute value (Panel (a)). Many dozens of principal components are required to span 80% of total variance (Panel (b))—though there is a non-trivial factor structure.[4] Thus, data mining not only uncovers notable

---

[4]For principal component analysis, we drop strategies with any missing values in the 1994-2020 sample, which may understate the diversity of the data.

Table 1: "Out-of-Sample" Returns from Mining Accounting Data

We sorts 29,000 data-mined strategies each June into 5 bins based on past 30-year mean returns ("in-sample") and computes the mean return over the next year within each bin ("out-of-sample"). Statistics are calculated by strategy, then averaged within bins, then averaged across sorting years. Decay is the percentage decrease in mean return out-of-sample relative to in-sample. We omit decay for bin 4 because the mean return in-sample is negligible. Data-mined returns are large and comparable to published returns, both in- and out-of-sample. Post-2004, out-of-sample returns are much weaker, consistent with information technology reducing mispricing.

	Panel (a): Out-of-Sample Returns 1994-2020							
In-	Equal-Weighted Long-Short Deciles				Value-Weighted Long-Short Deciles			
Sample	Past 30 Years (IS)		Next Year (OOS)		Past 30 Years (IS)		Next Year (OOS)	
Bin	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)
1	-59.3	-4.24	-49.4	16.7	-37.6	-2.06	-16.3	56.6
2	-29.1	-2.46	-18.9	35.1	-15.7	-1.02	-5.6	64.0
3	-13.3	-1.20	-3.2	75.9	-4.9	-0.33	-1.8	62.7
4	-0.3	-0.04	5.6		5.4	0.35	-0.0	
5	23.4	1.46	17.1	26.9	27.1	1.37	10.8	60.3

	Panel (b): Out-of-Sample Returns 2004-2020							
In-	Equal-Weighted Long-Short Deciles				Value-Weighted Long-Short Deciles			
Sample	Past 30 Years (IS)		Next Year (OOS)		Past 30 Years (IS)		Next Year (OOS)	
Bin	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)	Return (bps pm)	t-stat	Return (bps pm)	Decay (%)
1	-59.9	-4.04	-27.1	54.8	-37.2	-1.88	-4.6	87.6
2	-28.4	-2.32	-10.6	62.7	-14.4	-0.90	-0.7	95.1
3	-11.6	-1.00	-0.1	99.1	-3.8	-0.25	-1.7	55.5
4	2.3	0.19	7.8		6.1	0.40	-3.3	
5	25.9	1.53	17.4	32.8	28.4	1.38	2.3	91.9

out-of-sample performance, but also generates a very large number of distinct strategies.

These results replicate and extend the findings of Yan and Zheng (2017), who show that data mining produces out-of-sample returns in conservative, split-sample tests. This replication is important because Harvey and Liu (2020) criticize Yan and Zheng's methods and claim to find that the FDR from data mining is approximately 100%. Our results support the validity of Yan and Zheng's methods and are difficult to reconcile with

Table 2: Correlation Structure of Data-Mined Strategies with t-stats >2.0

We characterize the covariance structure of data-mined strategies using pairwise correlations (Panel (a)) or principal component analysis (Panel (b)). Both panels consider data-mined strategies with t-statistics greater than 2 in at least 10% of the 30-year in-sample periods from Table 1. The covariance structure is measured over the 1994-2020 sample. Each (Pearson) correlation coefficient is computed over all pairwise-complete observations. The figures report quantiles across correlation coefficients (eg. Q1 is the first percentile). Principal component analysis uses strategies with no missing values in the 1994-2020 sample. 80% of pairwise correlations are below .29 in absolute value. Many dozens of principal components are required to fully characterize the data.

Panel (a): Pairwise correlations									
Quantiles	Q1	Q5	Q10	Q25	Q50	Q75	Q90	Q95	Q99
Equal-Weighted	-0.42	-0.23	-0.15	-0.04	0.05	0.16	0.29	0.38	0.56
Value-Weighted	-0.35	-0.20	-0.13	-0.05	0.04	0.14	0.25	0.32	0.51

Panel (b): PCA Explained Variance (%)											
Number of PCs	1	5	10	20	30	40	50	60	70	80	90	100
Equal-Weighted	24	47	55	63	68	72	75	78	80	82	84	85
Value-Weighted	24	44	52	62	68	72	76	79	81	83	85	87

Harvey and Liu's claim that data mining produces approximately zero true discoveries.

2.3   Post-Sample Returns: Research vs Data Mining

We can now answer the question posed on page 1. How much does peer-reviewed research help predict cross-sectional returns compared to data mining?

To answer this question, we construct data-mined benchmarks for the published predictors. The data-mined benchmarks are found by searching the 29,000 accounting ratios for long-short strategy t-stats > 2.0 in the original sample periods. The benchmarks are selected to match their published counterparts in terms of equal- or value-weighting, and are signed to have positive mean returns in the original sample periods, before the t-stat > 2.0 filter is applied.

Figure 1 shows the result. It plots the mean returns of each class of predictor in event time, where the event is the end of the original sample periods. All strategies are normalized to have 100 bps mean return in the original samples. The figure then averages across strategies within each event-time month and then takes the trailing 5-year average to smooth out noise.

Post sample, peer-reviewed (solid) and data-mined (long-dash) predictors perform similarly. In fact, research and data mining lead to eerily similar event-time returns, with the data-mined returns resembling a Kalman-filtered version of the research returns. In this sense, it is difficult to reject the null that the research process is built off of data mining—at least for the research covered in the Chen and Zimmermann (2022) meta-study.

More positively, these results show that data-driven methods can generate substantial out-of-sample returns. Just back-testing accounting signals for t-stats > 2.0 leads to the same out-of-sample returns as drawing on the ideas from the best finance departments in the world. These results illustrate the potential of more sophisticated big data methods, like machine learning.

2.4 Even More Naive Data Mining Methods

Our data mining process (Section 2.1.1) just searches accounting ratios for t-stats > 2.0. But one can think of even more naive methods. How naive can one be and still generate research-like out-of-sample returns?
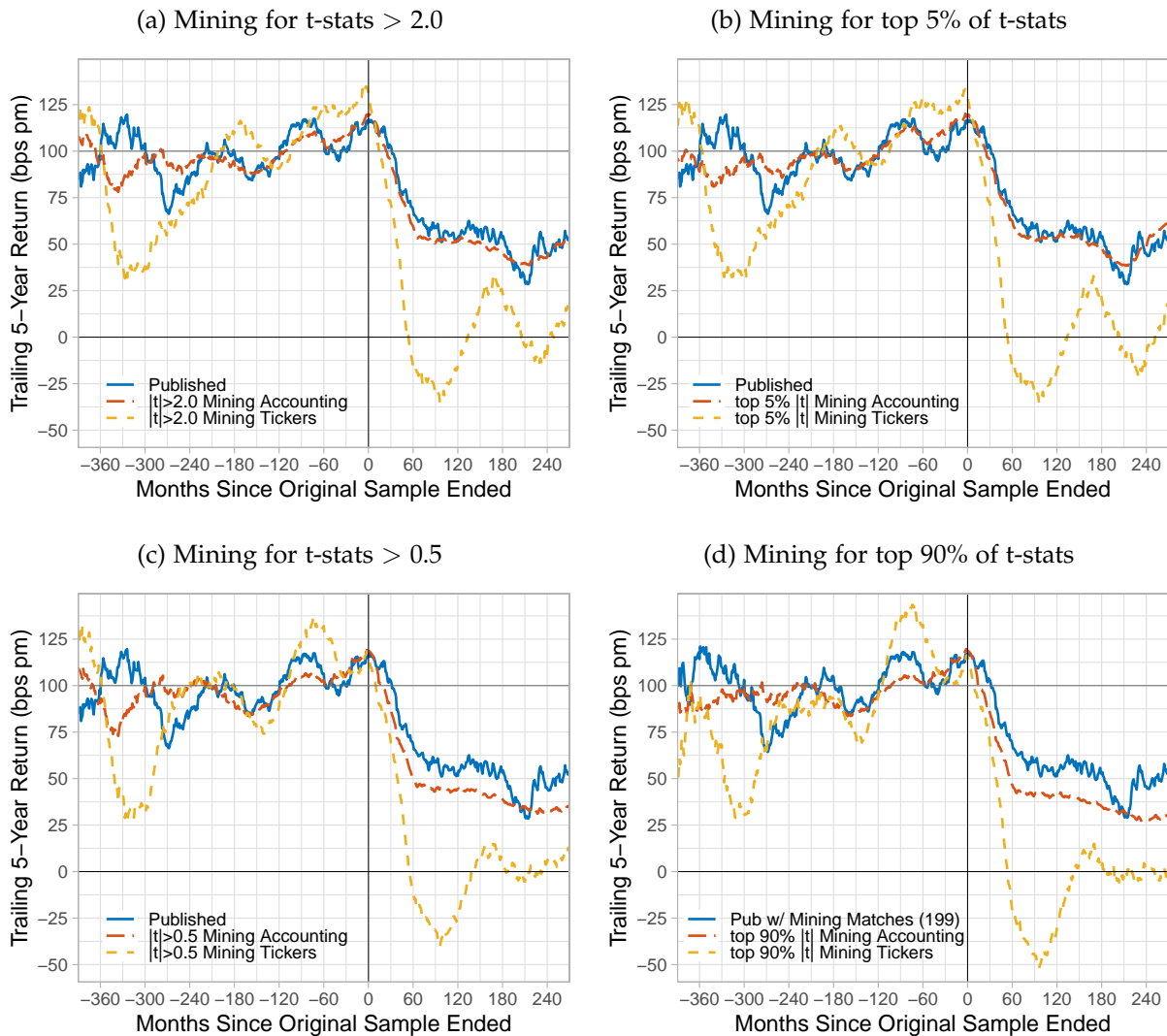
To answer this question, we examine an alternative data mining method proposed in Harvey (2017). Harvey asks his research assistant to "form portfolios based on the first, second, and third letters of the ticker symbol," leading to 3,160 long-short portfolios. We interpret his instructions as follows: Generate 26 portfolios by going long all stocks with a first ticker letter of "A," "B," "C," ..., and "Z." Generate 26 portfolios by doing the same for the second ticker letter, and add a 27th portfolio for tickers that have no second ticker letter. Apply the same to the third ticker. This process results in $26 + 27 + 27 = 80$ long portfolios. Finally, form $\binom{80}{2} = 3,160$ long-short portfolios by selecting all distinct pairs of the 80 long portfolios.

Figure 2 compares this ticker mining procedure to our baseline mining of accounting ratios. Panel (a) applies the same selection procedure as in Figure 1: we screen for t-stats > 2.0 in the original sample periods. Unlike mining accounting ratios, which leads to research-like post-sample returns, mining tickers leads to post-sample returns that are on average zero (yellow, short-dash line). Thus, mining tickers is too naive.

Panel (b) applies an alternative selection procedure: we screen for the top 5% of t-stats in the original sample periods. Mining accounting data still leads to very similar returns as research—in fact, the top 5% screen leads to returns that are even closer to research than the t-stat > 2.0 screen. This result suggests that research does not just screen for statistical significance, it actually focuses on the strongest signals available in the data.

Figure 2: Even More Naive Data Mining Methods

We compare published predictors (solid) to benchmarks made from data mining accounting ratios (long-dash) or tickers (short-dash). Benchmarks screen for a minimum t-stat (Panels (a) and (c)) or for the top X% of data-mined t-stats (Panels (b) and (d)) in the published papers' original sample periods. The plot shows the long-short returns in event time, where the event is the end of the original sample periods. Each predictor is normalized so that its mean original-sample return is 100 bps per month. Returns are averaged across predictors (by group) within each event month and then the trailing 5-year average is taken for readability. Extremely naive data mining generates post-sample returns in the ballpark of research, though mining tickers is too naive.

(a) Mining for t-stats > 2.0

(b) Mining for top 5% of t-stats

(c) Mining for t-stats > 0.5

(d) Mining for top 90% of t-stats

Mining tickers for the top 5% of t-stats still leads to post-sample returns of around zero, as in Panel (a).

The bottom panels examine much more lenient statistical screens. Panel (c) screens

11

for t-stats > 0.5 and Panel (d) screens for the top 90% of t-stats (buy all except for the worst 10% of t-stats). Since we sign all strategies to have positive original sample returns, these screens effectively take on only sign information (buy stocks that had positive mean returns in the past). The result is underperformance relative to research, but the difference is moderate. Mining accounting data still leads to post-sample returns that are about 75% as large as those from research.

Figure 2 illustrates two lessons about data mining. The first is that the data being mined is important. Some data, like accounting ratios, are very meaningful. Mining such data generates out-of-sample returns, even if the mining process uses only the sign of past mean returns. Other data, like tickers, are entirely uninformative. Mining these data only uncovers noise, and returns that vanish out-of-sample.

The second lesson is that more data mining does not necessarily mean worse out-of-sample performance. The accounting dataset is almost 10 times as large as the ticker dataset, yet it produces much stronger post-sample returns. In fact, we show that the amount of mining is actually irrelevant in our model of the "Cochrane Conjecture" (Section 3.4).

In summary, naively mining accounting data leads to post sample returns that are remarkably similar to those from the peer review process. While this result is negative for peer review, it illustrates the promise of big data methods. These methods can identify true predictability, even when applied in the most naive ways.

3   Does Risk-Based Theory Improve Outperformance?

Perhaps research that focuses on risk-based, equilibrium forces can find more stable returns, and thus outperform data mining (Cochrane (2009)). Put another way, perhaps the similar performance in Figure 1 is due to the fact that research publicizes disequilibrium mispricing. This mispricing then decays, as arbitrageurs push the economy toward equilibrium (McLean and Pontiff (2016)).

To examine this issue, we categorize predictors as risk-based, mispricing-based, or agnostic using the texts in the original papers (Section 3.1). We then examine the post-sample performance by category (Sections 3.2). We also examine categories based on the rigor of the theoretical justifications (Section 3.3). We close this section with a meta-theory for why theory may fail to help predict returns (Section 3.4).

3.1   Risk or Mispricing? According to Peer Review

We read each paper in the Chen and Zimmermann (2022) dataset and identify a passage of text that summarizes the main argument. These passages are typically taken from either the abstract, introduction, or conclusion. We then categorize each argument as "risk," "mispricing," or "agnostic." Each predictor was reviewed by two of the authors to prevent errors. We post all passages and categories on our Github repo, where anyone can review and comment on our categorizations (see DataInput/SignalsTheoryChecked.csv).

In a handful of cases, the text argues for liquidity explanations. We categorize these predictors as mispricing if the argument focuses on stock-specific measures of liquidity (Amihud (2002)) and risk if the argument focuses on a market-wide component (Pástor and Stambaugh (2003)). This method gives the risk category the best chance at finding post-sample returns, since idiosyncratic liquidity should improve over time (Chordia, Subrahmanyam, and Tong (2014)). Nevertheless, this issue affects only seven predictors, and has little impact on our main results.

Table 3 provides representative passages for each category. The risk and mispricing passages are straightforward: risk passages discuss risk, equilibrium, or market efficiency, while mispricing passages discuss mispricing or investor errors. Agnostic passages are slightly more difficult. Agnostic predictors are easy to classify when they claim agnosticism or provide arguments for both risk and mispricing. But in some cases, agnostic papers avoid discussing the explanation for predictability, and instead focus on the empirics (e.g. Boudoukh et al. (2007)).

Our analysis finds a remarkable consensus about the origins of cross-sectional predictability. This consensus is seen in Table 4, which counts the number of predictors in each theory category. Only 18% of cross-sectional predictors are judged by the peer review process to be due to risk. In contrast, 59% of predictors are due to mispricing. The remaining 23% of predictors are agnostic. Appendix Table A.2 shows that finance journals more commonly find risk explanations compared to accounting journals, but finance journals overall still attribute a minority of predictors to risk.

As a check on our manual classifications, we use software to count the ratio of "risk words" to "mispricing words" in each paper. For example, we count "utility," "maximize," and "priced" as risk words, and "behavioral," "optimistic," and "sentiment" as mispricing words (see Appendix A.1 for a full list). Table 4 shows order statistics of this ratio within each manually-classified theory category. The median ratio for risk-based predictors is 3.41—that is, risk words appear 3.4 times more frequently than mispricing words. Mirroring this result, mispricing-based predictors have a median ratio of 0.22, indicating five times as many mispricing words. Overall, this simple word count supports our

Table 3: Peer-Reviewed Risk and Mispricing Examples

Examples illustrate how we manually categorize predictors as risk, mispricing, or agnostic. Risk-to-mispricing words are counted by software and defined in Appendix A.1.

Reference	Predictor	Example Text	Risk to Mispricing Words
Panel (a): Risk			
Tuzel 2010	Real estate holdings	Firms with high real estate holdings are more vulnerable to bad productivity shocks and hence are riskier and have higher expected returns.	17.60
Bazdresch, Belo, and Lin 2014	Employment growth	We interpret this difference in average returns, which we refer to as the hiring return spread, as reflecting the relatively lower risk of the firms with higher hiring rates	7.32
Fama and MacBeth 1973	CAPM beta	The pricing of common stocks reflects the attempts of risk-averse investors to hold portfolios that are "efficient" in terms of expected value and dispersion of return.	2.31
Panel (b): Mispricing			
Ikenberry, Lakonishok, Vermaelen 1995	Share repurchases	Thus, at least with respect to value stocks, the market errs in its initial response and appears to ignore much of the information conveyed through repurchase announcements	0.05
Eberhart, Maxwell, and Siddique 2004	Unexpected R&D increase	We find consistent evidence of a mis-reaction, as manifested in the significantly positive abnormal stock returns that our sample firms' shareholders experience following these increases.	0.05
Desai, Rajgopal, Venkatachalam 2004	Operating Cash flows to price	CFO/P is a powerful and comprehensive measure that subsumes the mispricing attributed to all the other value-glamour proxies.	0.05
Panel (c): Agnostic			
Banz 1981	Size	To summarize, the size effect exists but it is not at all clear why it exists.	1.93
Boudoukh et al. 2007	Net Payout Yield	We show that the apparent demise of dividend yields as a predictor is due more to mismeasurement than alternative explanations such as spurious correlation, learning, etc.	1.00
Chordia, Subra, Anshuman 2001	Volume Variance	However, our findings do not lend themselves to an obvious explanation, so that further investigation of our results would appear to be a reasonable topic for future research.	0.21

manual categorizations. The distribution of risk to mispricing words for risk-based predictors is far to the right of the other categories.

Table 4: Risk or Mispricing? According to Peer Review

"Risk," "mispricing," or "agnostic" are constructed by manually reading papers and identifying key passages (Table 3). "Risk Words to Mispricing Words" shows the ratio of word counts in the papers (word list is in Appendix A.1). p05, p50, and p95 are the 5th, 50th, and 95th percentiles within each theory category. The list of all predictors, key passages, and classifications is available on our Github repo (DataInput/SignalsTheoryChecked.csv). Peer review attributes only 18% of predictors to risk.

Source of Predictability	Num Published Predictors			Risk Words to Mispricing Words		
	Total	1981-2004	2005-2016	p05	p50	p95
Risk	36	5	31	0.33	3.41	12.74
Mispricing	117	48	69	0.07	0.22	1.17
Agnostic	46	16	30	0.12	0.54	3.91
Any	199	69	130	0.07	0.33	7.02

The word counts also support our finding that risk explains a small minority of predictors. Across all papers, the median risk-to-mispricing word ratio is 0.33, meaning that mispricing-related words are typically mentioned 3 times as frequently as risk-related words.

The consensus in Table 4 is perhaps surprising given the tone in recent reviews on empirical cross-sectional asset pricing (e.g. Bali, Engle, and Murray (2016) and Zaffaroni and Zhou (2022)). These reviews provide a largely agnostic description of the origins of predictability, suggesting that peer review has come to a divided view, or that this topic has been too contentious to be available for open debate. Our results show that the literature favors mispricing, and that only a small minority of predictors are due to risk, as judged by the peer review process.

3.2 Post-Sample Returns of Risk vs Mispricing

Figure 3 shows the post-sample returns of risk-based, mispricing-based, and agnostic predictors. As in the previous figures, we plot trailing 5-year long-short returns in event time, and normalize each strategy to have 100 bps mean return in the original samples.

All three kinds of predictability decay by roughly 50% post-sample. So focusing on risk-based sources of predictability, as judged by peer review, does *not* lead to more stable out-of-sample returns. Indeed, risk-based predictability seems to decay even more than mispricing-based predictability. This result is robust to adjusting for CAPM exposure (Appendix Figure A.2), though we focus on raw returns because the CAPM often holds

Figure 3: Post-Sample Returns by Peer-Reviewed Explanation

The plot shows the mean long-short returns of published predictors in event time, where the event is the end of the original sample periods. Each predictor is normalized so that its mean in-sample return is 100 bps per month. Predictors are grouped by theory category based on the arguments in the original papers (Table 3). We average returns across predictors within each month and then take the trailing 5-year average for readability. For all categories of theory, predictability decays by roughly 50% post-sample. If anything, risk-based predictors decay more than other predictors.
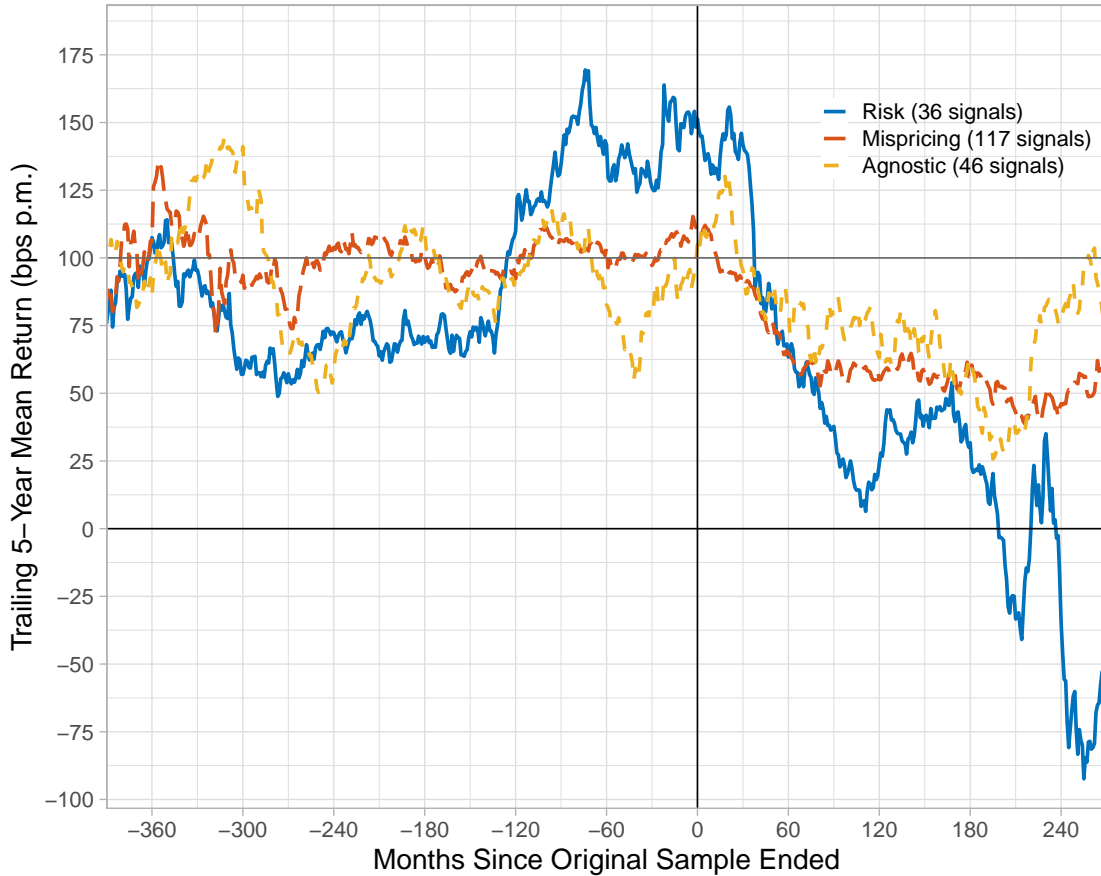
in risk-based models of cross-sectional predictability (e.g. Zhang (2005); Tuzel (2010)). The underperformance in Figure 3 comes from just 36 risk-based predictors, which raises questions about statistical significance.

Table 5 examines statistical significance in a regression framework (following McLean and Pontiff (2016)). Specification (1) regresses monthly long-short returns on a post-sample indicator and its interaction with an indicator for risk-based predictors. Returns are normalized to be 100 bps per month in-sample, so the post-sample coefficient implies

16

that returns decay by 42 percent overall (across all types of predictors).[5] The interaction coefficient implies that risk-based predictors have an additional decay of 29 percentage points, for a total decay of 71 percent. The additional decay of risk predictors is only marginally significant, highlighting the limitations of the sample size.

Table 5: Regression Estimates of Risk vs Mispricing Effects on Predictability Decay

We regress monthly long-short returns on indicator variables to quantify the effects of peer-reviewed risk vs mispricing explanations on predictability decay. Each strategy is normalized to have 100 bps per month returns in the original sample. "Post-Sample" is 1 if the month occurs after the predictor's sample ends and is zero otherwise. "Post-Pub" is defined similarly. "Risk" is 1 if peer review argues for a risk-based explanation (Table 3) and 0 otherwise. "Mispricing" and "Post-2004" are defined similarly. Parentheses show standard errors clustered by month. "Null: Risk No Decay" shows the p-value that tests whether risk-based returns do not decrease post-sample ((1) and (3)) or post-publication ((2) and (4)). The decay in risk-based predictors is highly statistically significant, and inconsistent with the hypothesis that risk theory uncovers stable expected returns.

	LHS: Long-Short Strategy Return (bps pm, scaled)				
RHS Variables	(1)	(2)	(3)	(4)	(5)
Intercept	100.0	100.0	100.0	100.0	102.4
	(6.4)	(6.4)	(6.4)	(6.4)	(6.8)
Post-Sample	-42.3	-25.3	-36.5	-24.4	0.7
	(8.6)	(11.7)	(10.3)	(15.3)	(14.5)
Post-Pub		-21.0		-14.9	
		(12.1)		(17.5)	
Post-Sample x Risk	-28.7	-18.5	-34.4	-19.5	-23.4
	(15.4)	(20.2)	(17.1)	(22.8)	(15.2)
Post-Pub x Risk		-14.2		-20.3	
		(27.2)		(30.2)	
Post-Sample x Mispricing			-8.1	-1.3	
			(7.8)	(15.5)	
Post-Pub x Mispricing				-8.7	
				(17.5)	
Post-2004					-59.6
					(16.6)
Null: Risk No Decay	< 0.1%	< 0.1%	< 0.1%	< 0.1%	< 0.1%

Nevertheless, there is plenty of data to show that risk-based explanations fail to

---

[5]This decay implies 58% of returns remains post-sample if each predictor-month is weighed equally. This is a bit higher than the 53% found on page 1, which weighs each month equally, and thus focuses more on returns further from the original sample.

prevent post-sample decay. This result is shown in the row "Null: Risk No Decay," which tests the hypothesis that the sum of the Post-Sample and Post-Sample $\times$ Risk coefficients is non-negative. The test rejects this hypothesis at the 0.1% level.

Specifications (2)-(4) show robustness. Specification (2) adds a post-publication indicator, specification (3) adds an indicator for mispricing explanations, and specification (4) adds both. All three alternative specifications arrive at risk-based predictors decaying by an additional 30 to 40 percentage points. Specification (4) implies that post-publication, being risk-based implies an additional 20 + 20 = 40 percentage points of decay, for a total decay of 24 + 15 + 40 = 79%.

Additional robustness is shown in specification (5), which controls for the idea that information technology has led to weaker predictability post-2004 (Chordia, Subrahmanyam, and Tong (2014)). In this specification, risk-based predictors still decay by an additional 23 percentage points.

A more refined control for time effects is found in Figure 4. As in Figure 1, we construct data-mined benchmarks by searching 29,000 accounting signals for t-stats > 2.0 in the risk-based predictors' original sample periods. These benchmark returns are thus exposed to the same time effects as the risk-based predictors, such as business cycles and interest rate regimes.

Controlling for time effects in this way, risk-based predictors still underperform. Their trailing returns (solid line) outperform in the three years after event month zero, but these returns are not out-of-sample: the 5-year window overlaps with the original sample period. Moreover, papers are not published until about 48 months after the end of the original samples. Trailing returns that are fully out of sample are largely below the data-mined benchmarks (dashed line). The analogous plots for mispricing-based and agnostic predictors are in Appendix Figure A.1, and show that agnostic predictors slightly outperform, an issue we return to in Section 5.1.

Taken together, these results demonstrate an important asymmetry in how investors learn from academic publications. If investors learn about mispricing, then they will buy undervalued stocks, sell overvalued stocks, and predictability will weaken post-sample (McLean and Pontiff (2016)). But this logic does not hold for risk-based publications. Indeed, the same logic suggests that, if investors learn about risk from publications, they will buy the safe stocks, sell the risky stocks, and predictability will *strengthen* post-sample.

Figures 3 and 4 show that this symmetry does not hold in the data. Instead, both risk-based and mispricing-based predictors decay post-sample. If there is any difference, it is risk-based predictors that decay more.

18

Figure 4: Risk-Based Predictors vs Data-Mined Benchmarks

'Published' includes only predictors that are based on risk according to peer review (Table 3). '$|t| > 2.0$ Mining Accounting' is a data-mined benchmark formed by filtering 29,000 strategies for $|t| > 2.0$ in published predictors' original sample periods. The plot shows long-short returns in event time, where the event is the end of the original sample periods. All predictor returns are normalized to average 100 bps in the original samples. Risk-based predictors (as judged by peer review) underperform data mined predictors that are exposed to the same time effects.
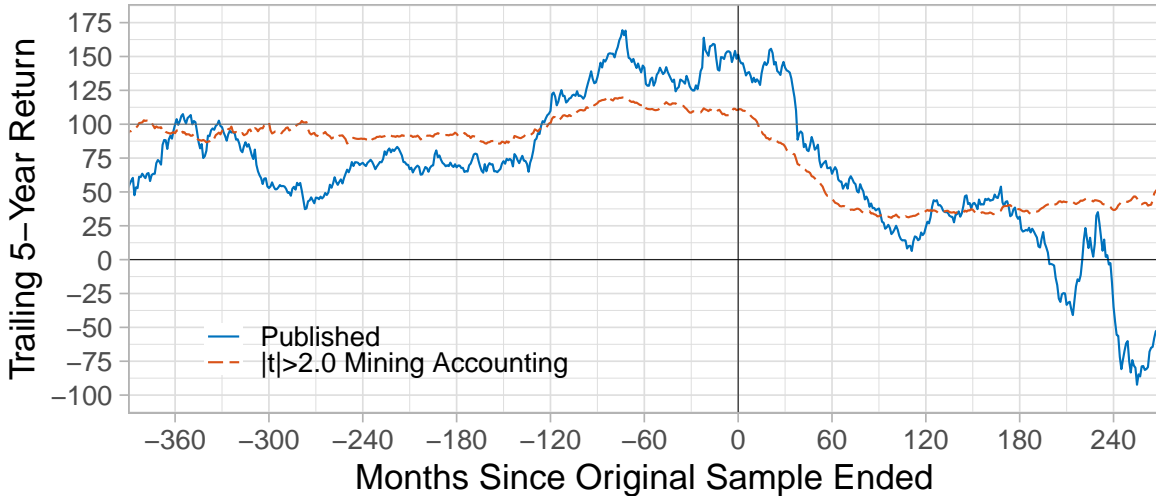
This broken symmetry implies that either peer review systematically mislabels mispricing as risk or that peer review on average finds unstable risk that decays over time. Neither implication is positive for the peer review process. Fortunately, peer review labels only a small minority of predictors as risk.

3.3 Do Mathematical Models Help?

A common belief is that theory protects against post-sample decay by restricting the number of possible signals (e.g. Harvey, Liu, and Zhu (2016)).[6] Perhaps the risk-based predictors in Section 3.2 are not restricted enough, since many of the risk-based predictors are supported by informal arguments rather than rigorous equilibrium theory. Does focusing on predictors supported by mathematical models help?

To address this question, we categorize predictors by the rigor of the mathematical model (if any) that is used as supportive evidence. The categories we consider are stylized model (e.g. a two-period model), dynamic equilibrium (many periods), or

---

[6]The meta-theory in Section 3.4 implies this belief is misguided.

quantitative equilibrium (calibrated to match key moments in the data). We then examine the post-sample performance by model type.

Figure 5 shows the result. 25 of the 199 predictors in our sample are supported by a mathematical model, and only 6 of these are quantitative equilibrium. While this sample is small, point estimates imply the opposite of what is commonly believed. The mean normalized post-sample return is monotonically *decreasing* in modeling rigor.

Figure 5: Post-Sample Returns by Model Rigor

Each marker represents one published predictor's post-sample mean return normalized by its original-sample return. 'Stylized,' 'dynamic,' and 'quantitative' are the type of models used as supporting evidence for the predictor, with 'quantitative' models being dynamic or asymmetric information models calibrated to match important moments in the data. Diamonds show means across predictor by category. The full reference for each acronym can be found at `https://github.com/OpenSourceAP/CrossSection/blob/master/SignalDoc.csv`. The estimated relationship between modeling rigor and post-sample performance is monotonic and negative.
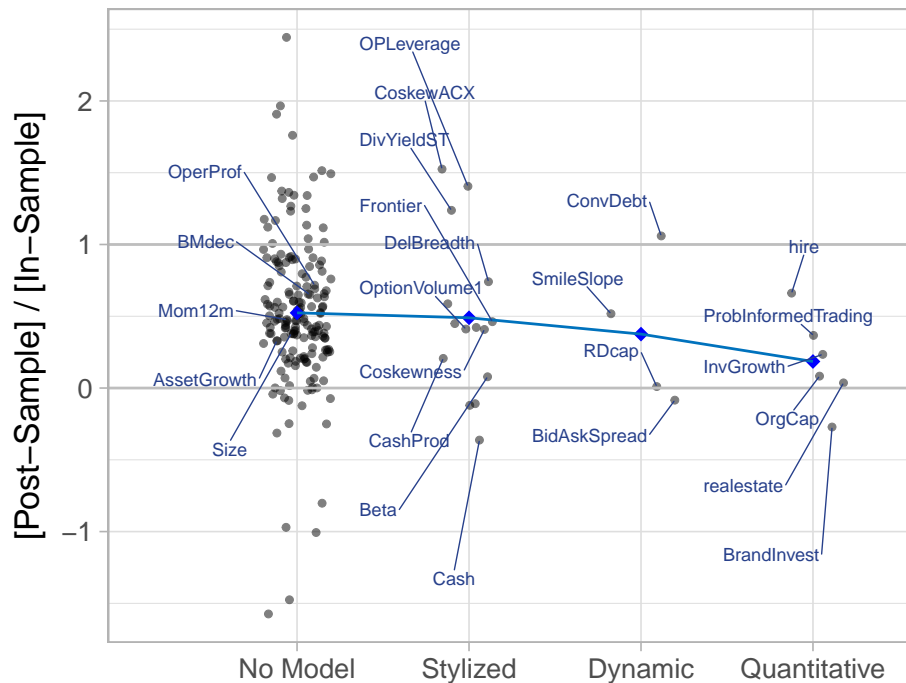
This result is notable given the impressive work behind quantitative equilibrium models. This work is not just theoretical: it is also computational and empirical. 'realestate' is based on Tuzel's (2010) general equilibrium production economy with heterogeneous firms. Despite the difficulties in solving this class of models, Tuzel does not simplify for tractability, and instead computes the non-linear equilibrium using Krusell and Smith

(1998) approximate aggregation and the parameterized expectations algorithm of Marcet (1991). She then calibrates the equilibrium to match many moments in the data. The calibration shows that her model is not just a qualitative description, but that it is a quantitative match for the U.S. economy. Despite all of this rigor, 'realestate' returned only 1 bps per month between the end of Tuzel's sample period (2005) and the end of our dataset (2022).

The other quantitative equilibrium models in Figure 5 are not general equilibrium, but they are still impressive. Each one solves a dynamic or asymmetric information model numerous times to ensure that the selected model matches many features of real world data. This work leads to predictors that under perform papers that use stylized models or informal arguments.

One interpretation is that a talented theorist can justify nearly any empirical pattern (the Sonnenschein-Mantel-Debreu "Anything Goes Theorem," Mas-Colell, Whinston, and Green (1995)), especially with the aid of modern computing power. If this is the case, then the support of a quantitative equilibrium model may not provide any additional information about the predictor. Another interpretation is that the set of theories is restricted (perhaps by the judgement of peer review) but that these restrictions are unrelated to the risks of concern to real-world investors. We provide a more formal statement of the latter interpretation in Section 3.4.

Regardless of the interpretation, these results raise questions about the idea that economic priors should be imposed in the peer review process (Harvey, Liu, and Zhu (2016); Harvey (2017)). At least some peer reviewers would have the prior that quantitative equilibrium models are a sign of out-of-sample robustness. Figure 5 suggests that such priors need updating.

3.4   A Model of the "Cochrane Conjecture"

Our results may be surprising given the common belief that theory protects against data mining bias. We call this belief the "Cochrane Conjecture," after the discussion of this issue in the Cochrane (2009) textbook. This section provides a formal analysis of this conjecture and provides sufficient conditions for its validity. The model is a slight generalization of models in Chen and Zimmermann (2020) and Chen and Velikov (2022).

The long-short return of predictor i in month $t = 1, 2, ...$ depends on whether t is in

the original sample ($t \leq T_i$) or post-sample ($T_i < t \leq T_i + T_i^{OOS}$):

$$r_{i,t} = \begin{cases} \mu_i + \varepsilon_{i,t} & t \leq T_i \\ \mu_i + \Delta\mu_i + \varepsilon_{i,t} & T_i < t \leq T_i + T_i^{OOS} \end{cases}. \tag{1}$$

where $\mu_i \equiv E(r_{i,t}|t \leq T_i)$ is the expected return in the original sample, $\Delta\mu_i \equiv E(r_{i,t}|T_i < t \leq T_i + T_i^{OOS}) - \mu_i$ is the change in expected returns change post-sample, and $\varepsilon_{i,t}$ is a zero mean residual. Sample mean returns during the original and post-sample periods are denoted by $\bar{r}_i \equiv T_i^{-1} \sum_{t=1}^{T} r_{i,t}$ and $\bar{r}_i^{OOS} \equiv (T_i^{OOS})^{-1} \sum_{t=T_i+1}^{T_i+T_i^{OOS}} r_{i,t}$, respectively. The post-sample decay is defined as

$$\text{Decay}_i \equiv \text{Sign}(\bar{r}_i)\left(\bar{r}_i - \bar{r}_i^{OOS}\right), \tag{2}$$

where $\text{Sign}(\bar{r}_i)$ flips the long and short legs if the original sample mean return is negative. These expressions are definitions rather than economic restrictions.

The only economic restriction is that post-sample residuals are unpredictable using original sample information:

$$E\left(\varepsilon_{i,t} \left| \mathcal{I}_i^{IS}, T_i < t \leq T_i + T_i^{OOS}\right.\right) = 0 \tag{3}$$

where $\mathcal{I}_i^{IS}$ represents any information available at month T_i for predictor i (e.g. the original sample mean return $\bar{r}_i$).

Data mining involves searching through a set $\mathcal{D}$ of predictors (e.g. the accounting-based predictors in Section 2.1.1) and selecting those with $|\bar{r}_i|$ above some threshold h. The expected post-sample decay from this process is

$$E\left(\text{Decay}_i \mid |\bar{r}_i| > h, i \in \mathcal{D}\right) = E\left(\text{Sign}(\bar{r}_i)(-\Delta\mu_i + \bar{\varepsilon}_i) \mid |\bar{r}_i| > h, i \in \mathcal{D}\right), \tag{4}$$

where post-sample residuals are missing due to Equation (3). However, the original sample's mean residual $\bar{\varepsilon}_i$ remains, due to the selection on $\bar{r}_i > h$. Thus, there are two reasons for post-sample decay: either the fundamental expected return shifts toward zero ($\text{Sign}(\bar{r}_i)\Delta\mu_i < 0$) or the selection process picks up lucky original sample reutrns ($\text{Sign}(\bar{r}_i)\bar{\varepsilon}_i > 0$).

Using economic theory amounts to searching through a different set $\mathcal{T}$ (e.g. risk-based predictors). The question is: under what conditions does searching through $\mathcal{T}$ lead to less decay? The following proposition provides sufficient conditions:

Proposition 1. *(Cochrane Conjecture) If the following two inequalities hold:*

$$E\left(Sign(\bar{r}_i)\mu_i \mid |\bar{r}_i|, i \in \mathcal{T}\right) > E\left(Sign(\bar{r}_i)\mu_i \mid |\bar{r}_i|, i \in \mathcal{D}\right) \tag{5}$$

$$E\left(Sign(\bar{r}_i)\Delta\mu_i \mid |\bar{r}_i|, i \in \mathcal{T}\right) > E\left(Sign(\bar{r}_i)\Delta\mu_i \mid |\bar{r}_i|, i \in \mathcal{D}\right), \tag{6}$$

then theory leads to less decay than data mining:

$$E\left[Decay_i \mid |\bar{r}_i| > h, i \in \mathcal{T}\right] < E\left[Decay_i \mid |\bar{r}_i| > h, i \in \mathcal{D}\right].$$

The proof is in Appendix A.3.

Proposition 1 provides a formal justification for the argument in Chapter 7 of Cochrane (2009). There, Cochrane describes the problem of "dredging up spuriously good in-sample pricing," and argues that "the best hope for finding pricing factors that are robust out of sample... ...is to try to understand the fundamental macroeconomic sources of risk."

As implied by the quote, improved robustness is not guaranteed, but it obtains under some conditions. According to Proposition 1, one condition is that understanding risk leads to higher expected returns than data mining—holding fixed summary statistics about past performance (Equation (5)). Another condition is that understanding risk leads to more stable expected returns—once again holding fixed summary statistics (Equation (6)). The importance of risk and equilibrium is implicit in these expressions. Unlike predictability that is based on mispricing, risk-based predictability should be stable both within the original sample ($\bar{r}_i$ is due to μ_i and not $\bar{\varepsilon}_i$) and post-sample ($\Delta\mu_i$ does not shift expected returns toward zero). If both of these conditions hold, then Proposition 1 says improved out of sample robustness *must* obtain.

Our empirical results imply that at least one of these conditions fails to hold for the understanding of risk embodied by the papers in the Chen and Zimmermann (2022) dataset. Either cause is consistent with a disconnect between the risks identified in these papers and the risks that are of concern to real-world investors. This disconnect is found in many surveys of real-world investors, from finance professionals (Mukhlynina and Nyborg (2020); Chinco, Hartzmark, and Sussman (2022)), to millionaires (Bender et al. (2022)), to tenured finance professors (Doran and Wright (2007)). In all of these surveys, risks that are important according to peer review are unimportant for real world decisions.

Notably, the size of the sets $\mathcal{T}$ and $\mathcal{D}$ does not appear in Proposition 1. In other words, the *amount* of data mining is irrelevant, in contrast the common intuition that more tests imply more false discoveries (e.g. Harvey, Liu, and Zhu (2016)). Instead, what matters is whether theory or data mining provides a better signal of the fundamental expected

returns (μ_i and $\Delta\mu_i$). This theory of data mining is consistent with our finding that data-mined ticker predictors decay more than accounting predictors (Figure 2)—despite the fact that the ticker-based predictors are mined from a much smaller dataset.

4 Data-Mined Alternatives to B/M and Momentum

The B/M and momentum predictors have captured the imagination of asset pricing scholars going back to Fama and French (1992) and Jegadeesh and Titman (1993) (see also Stattman (1980)). The previous sections suggest that these predictors are not as special as previously thought. This section takes a closer look at this possibility by constructing data-mined alternatives to B/M and momentum.

4.1 Data-Mined Alternatives to Fama and French's (1992) B/M

Suppose that it is the year 1992 and you have just read Fama and French (1992). But instead of using Fama and French's insights, you decide to data mine for accounting ratios with similar predictive power to B/M. What kinds of predictors would you find? What would your post-sample returns be?

Table 6 answers these questions. It lists 20 of the 171 predictors with long-short decile returns within 30% of B/M's and t-stats within 10% of B/M's, using the original 1963-1990 sample. The predictors are ranked by their similarity with B/M in terms of mean returns.

The data-mined predictors include themes that have been found in the cross-sectional literature. Notably, these themes include many variables that were document *after* Fama and French (1992). The table includes measures of accruals (Sloan (1996)), profitability (Fama and French (2006)), asset growth (Cooper, Gulen, and Schill (2008)), and equity issuance (Pontiff and Woodgate (2008)). Thus, data mining works, in part, by uncovering the same ideas as the peer-review process.

The bottom of Table 6 shows how these data mined predictors perform. These 171 predictors earned on average 83 bps per month in the 1963-1990 sample, a touch below the 96 bps per month earned by B/M. Post 1991, the 171 predictors earned on average 69 bps per month, outperforming B/M by 7 bps. This result shows that the data mining not only performs as well as peer reviewed research overall, but also performs as well as the most famous predictor in the literature, with arguably the most well-studied risk-based foundations (Fama and French (1993); Zhang (2005); Chen (2018)).

Table 6: 20 Data-Mined Predictors With Returns Similar to Fama-French's B/M (1992)

Table lists 20 of the 171 data-mined signals that performed similarly to Fama and French's (1992) B/M in the original 1963-1990 sample period in terms of mean returns and t-stats. Signals are ranked according to the absolute difference in mean in-sample return. Sign = -1 indicates that a high signal implies a lower mean return in-sample. Data mining picks up themes found by peer-reviewed research (e.g. investment, equity issuance, accruals) and leads to similar out-of-sample performance as Fama and French's B/M.

Similarity Rank	Signal	Sign	Mean Return (% p.m.) 1963-1990	1991-2021
Peer-Reviewed				
	Book / Market (Fama-French 1992)	1	0.96	0.62
Data-Mined				
1	Δ[Assets]/lag[Operating expenses]	-1	0.96	0.90
2	Δ[PPE net]/lag[Sales]	-1	0.96	0.80
3	[Market equity FYE]/[Depreciation & amort]	-1	0.95	0.66
4	Δ[Assets]/lag[Cost of goods sold]	-1	0.95	0.86
5	Δ[Assets]/lag[SG&A]	-1	0.95	0.82
6	Δ[PPE net]/lag[Gross profit]	-1	0.98	0.51
7	Δ[PPE net]/lag[Current liabilities]	-1	0.94	0.90
8	[Depreciation (CF acct)]/[Capex PPE sch V]	1	0.97	0.78
9	[Market equity FYE]/[Depreciation depl amort]	-1	0.94	1.03
10	[Stock issuance]/[Capex PPE sch V]	-1	0.94	0.93
	. . .			
101	[Market equity FYE]/[Current assets]	-1	1.15	0.89
102	[Market equity FYE]/[Common equity]	-1	1.14	0.51
103	Δ[Cost of goods sold]/lag[Cost of goods sold]	-1	0.75	0.94
104	Δ[Receivables]/lag[Invested capital]	-1	0.76	0.46
105	Δ[Receivables]/lag[PPE net]	-1	0.76	0.46
	. . .			
167	Δ[Assets]/lag[IB adjusted for common s]	-1	0.67	-0.02
168	Δ[Cost of goods sold]/lag[Current liabilities]	-1	0.67	0.68
169	Δ[Long-term debt]/lag[Operating expenses]	-1	0.67	0.50
170	Δ[Depreciation & amort]/lag[Invested capital]	-1	0.67	0.68
171	Δ[Current liabilities]/lag[Inventories]	-1	0.67	0.49
	Mean Data-Mined		0.83	0.69

Table 7: 20 Data-Mined Predictors That Perform Similarly to Jegadeesh and Titman's Momentum (1993)

Table lists 20 of the 44 data-mined signals that performed similarly to Jegadeesh and Titman's (1993) 12-month momentum in the original 1964-1989 sample period in terms of mean returns and t-stats. Signals are ranked according to the absolute difference in mean in-sample return. Sign = -1 indicates that a high signal implies a lower mean return in-sample. Data mining picks up themes found by peer-reviewed research (e.g. profitabitability, investment) and leads to similar out-of-sample performance as Jegedeesh and Titman's momentum.

Similarity Rank	Signal	Sign	Mean Return (% p.m.)	
			1964-1989	1990-2021
Peer-Reviewed				
	12-Month Momentum (Jegadeesh-Titman 1993)	1	1.38	0.66
Data-Mined				
1	[Retained earnings unadj]/[Market equity FYE]	1	1.38	-0.19
2	[Retained earnings unadj]/[Assets other sundry]	1	1.40	0.15
3	[PPE and machinery]/[Current liabilities]	1	1.42	0.38
4	[Retained earnings unadj]/[Cash & ST investments]	1	1.42	0.17
5	[PPE and machinery]/[Capital expenditure]	1	1.50	0.79
6	[Retained earnings unadj]/[Invest & advances other]	1	1.51	0.03
7	[Investing activities oth]/[Nonoperating income]	1	1.52	0.10
8	[Income taxes paid]/[PPE net]	1	1.22	0.14
9	[PPE and machinery]/[Capex PPE sch V]	1	1.56	0.80
10	[Market equity FYE]/[Current assets]	-1	1.20	0.88
	. . .			
21	[Depreciation (CF acct)]/[Market equity FYE]	1	1.10	0.72
22	[Retained earnings unadj]/[Common equity]	1	1.66	-0.08
23	Δ[Assets]/lag[Current assets]	-1	1.09	1.26
24	Δ[PPE (gross)]/lag[Operating expenses]	-1	1.09	0.70
25	[Funds from operations]/[Market equity FYE]	1	1.07	-3.51
	. . .			
40	[Deprec end bal (Sch VI)]/[Market equity FYE]	1	1.02	1.03
41	[Market equity FYE]/[Cost of goods sold]	-1	1.00	0.70
42	[Rental expense]/[Market equity FYE]	1	1.00	0.84
43	Δ[Invested capital]/lag[Current assets]	-1	0.97	1.32
44	[Retained earnings unadj]/[Receiv current other]	1	1.79	0.19
	Mean Data-Mined		1.29	0.48

4.2 Data-Mined Alternatives to Jegadeesh and Titman's (1993) Momentum

Table 7 applies the same exercise to Jegadeesh and Titman's (1993) 12-month momentum. Here, peer review outperforms, with momentum earning 66 bps per month post-sample

compared to 48 bps for data mining. This outperformance, however, may be related to the fact that 12-month momentum is not part of our dataset of accounting ratios (see Section 2.1.1). Chen and Dim (2023) show that data mining can reliably uncover predictability from past returns data, though their technique requires more sophisticated empirical Bayes methods.

Many of the themes seen in Table 6 show up again in Table 7, though we see some unusual variables like rental expense, depreciation, and income taxes. The Appendix lists predictors related to Banz's (1981) Size predictor (Table A.1), which also include well-known themes (investment and profitability) and more unusual variables (investment tax credits and interest expense).

Overall, these results illustrate the potential of data mining for helping us understand the economics of predictability. One may have thought that linking investment or profitability to expected returns requires Ph.D.-level insight. But it turns out that data mining based on basic accounting principles can systematically uncover these patterns.

5 Robustness

This section demonstrates robust. Section 5.1 shows robustness to matching closely on original sample statistics and excluding correlated strategies. Section 5.2 examines a continous measure of risk based on peer-reviewed text. Section 5.3 shows robustness to using factor model measures of risk.

5.1 Closely-Matched Original Sample Statistics and Excluding Correlated Returns

The previous results may be unfair to peer-reviewed predictors, as the data-mined predictors may have stronger statistical support. It would not be fair, for example, to compare data-mined predictors with t-statistics of 6.0 a peer-reviewed predictor with a t-statistic of 2.0.

To control for this issue, we use a more restrictive matching procedure. For each published predictor, we match with data-mined predictors that have t-statistics within 10% and mean returns within 30% of the published predictors, using the original sample periods. As in Section 2, we also restrict the data-mined predictors to match the published ones in terms of equal- or value-weighting. We then repeat our primary post-sample tests (Figures 1 and 4).

Figure 6 shows the result. Data mining continues to perform similarly to peer review overall (Panel (a)) and outperforms risk-based predictors out-of-sample (Panel (b)). This results is perhaps natural, as it is unlikely that data mining tends to uncover larger t-statistics than peer review, since researchers themselves can use data mining (Chen (2024)).

Panels (c) and (d) look closer at mispricing and agnostic predictors. Data mining closely mimics the returns of data-mined predictors but it somewhat underperforms agnostic predictors. These results are consistent with data-mining benchmarks that simply screen on t-stats > 2.0 (Appendix Figure A.1), so it is not the more restricted screen that generates the outperformance of agnostic predictors. A potential explanation for this outperformance is that many of the agnostic predictors are based on past returns. Past-return predictors are missing from the accounting ratios we use for the data mining benchmarks (Section 2.1.1). Several agnostic past-return predictors have performed quite well post-sample (e.g. Moskowitz and Grinblatt's industry momentum (1999) and several versions of Heston and Sadka's seasonal momentum (2008)).

Another potential concern is that correlations may be driving our results. This idea can be motivated by the idea that expected returns are driven by correlations with risk factors.

To address this concern, we additionally exclude data-mined strategies that have correlations of more than 0.10 with the target published strategy in the original samples. This additional filter (red long-dashed line) has little effect on the path of 5-year returns in any of the four panels of Figure 6.

This robustness is natural given the diversity of data-mined predictors from Table 2. The majority of data-mined predictors with t-stats > 2.0 have correlations less than 0.25 in absolute value.

5.2   A Continuous Measure of Risk

The previous measures of risk are binary: a predictor is either risk-based or not. But predictors may be due to a mixture of risk and mispricing. Perhaps a more continuous measure of risk could help predict post-sample returns.

To examine this possibility, Figure 7 plots post-sample returns against the ratio of risk words to mispricing words in the published papers. The risk and mispricing words are defined as in Table 4 (see also Appendix A.1).

The figure shows a negative relationship between post-sample returns and the ratio of risk to mispricing words. Thus, the underperformance of risk predictors is not due

Figure 6: Data-Mining vs Peer-Review Excluding Correlated Returns

We match data-mined with published strategies based on original-sample t-stats (as in Figures 1 and 4) but now we drop data-mined strategies if they have t-stats that differ by more than 10% or mean returns that differ by more than 30% (yellow short dash). We additionally drop data-mined strategies that are more than 10% correlated with published strategies in the original sample (red long dash). The similarity in post-sample returns is not driven by correlations.

(a) All

(b) Risk-based

(c) Mispricing-based

(d) Agnostic

to an artifical binary classification. This result is also consistent with the monotonically negative relationship between model rigor and post-sample performance (Figure 5).

5.3 Factor Model Measures of Risk

Factor models are commonly used to measure risk in the asset pricing literature. Instead of using peer-reviewed text, this section uses this more traditional method of measuring risk.

Figure 7: Post-Sample Returns vs Risk to Mispricing Words

Each marker represents one published predictor's mean return. The regression line is fitted with OLS. The full reference for each acronym can be found at `https://github.com/OpenSourceAP/CrossSection/blob/master/SignalDoc.csv`. The relationship between risk words and post-sample returns is negative.
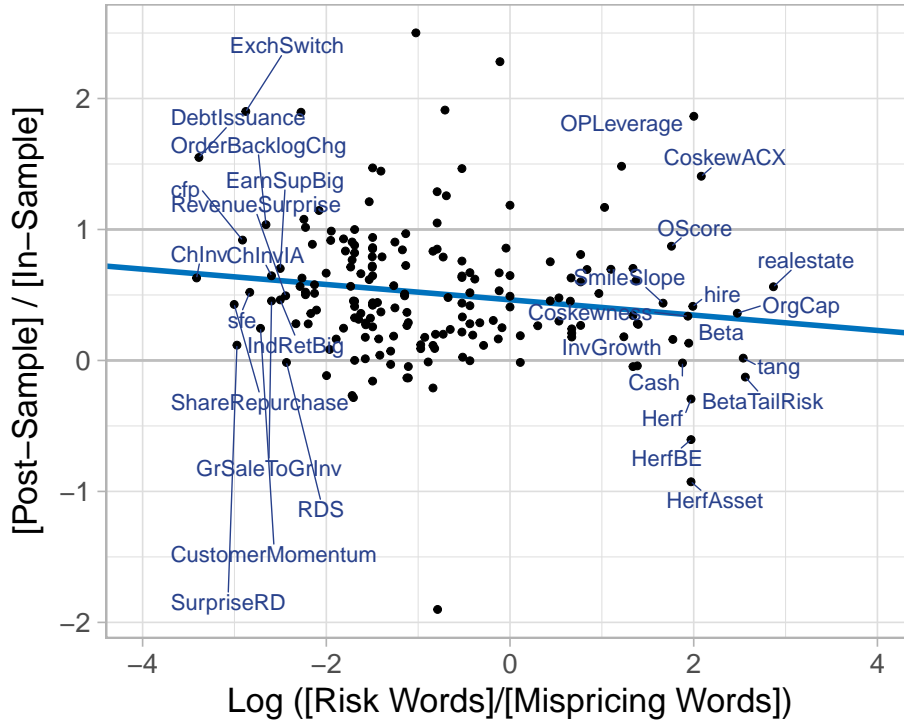
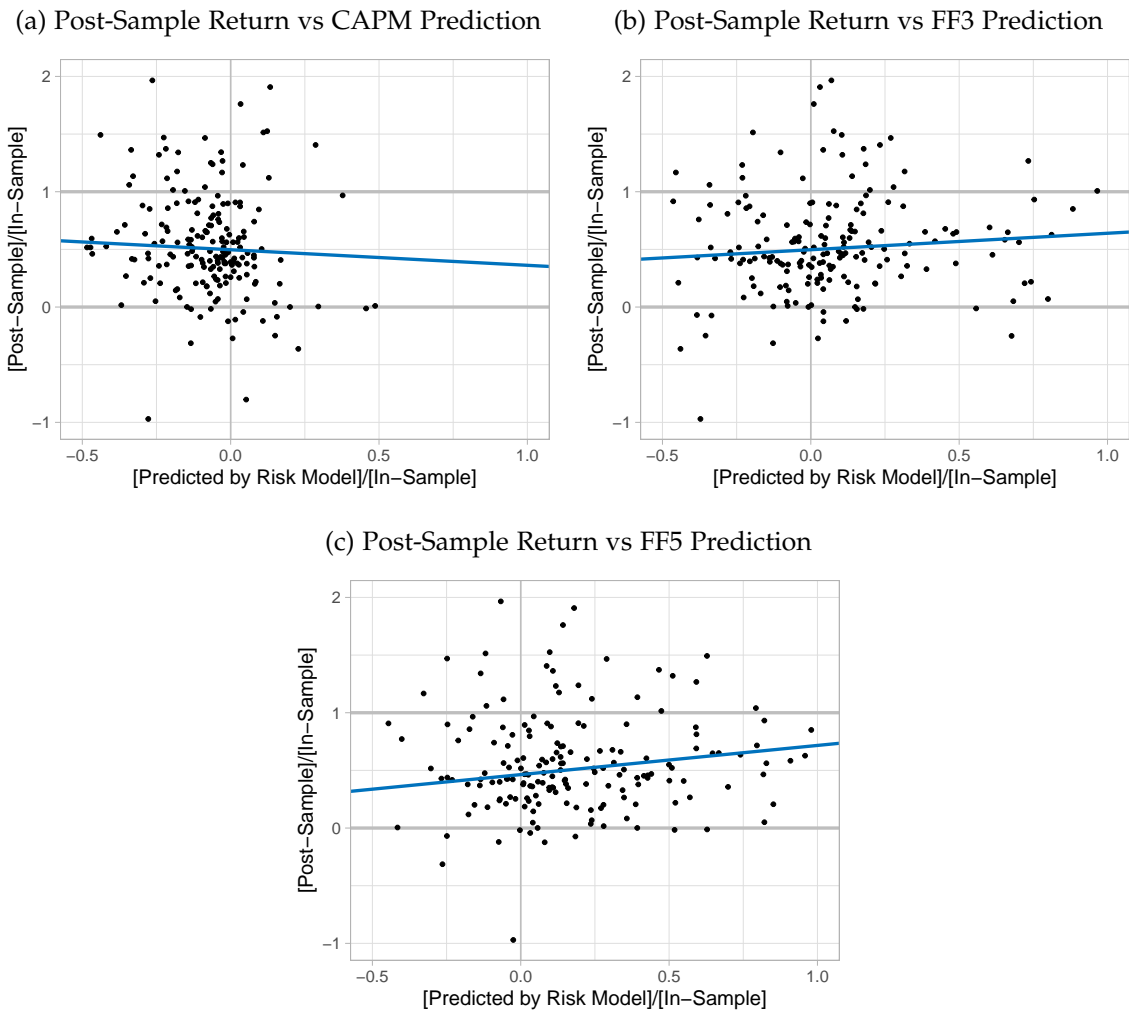In particular, we measure risk using the CAPM, Fama-French 3 (FF3), and Fama-French 5 (FF5) factor models.

For each published long-short portfolio i, we estimate exposure to factor k $(\hat{\beta}_{i,k})$ using time-series regressions on the original papers' sample periods. According to the factor models, the estimated expected return is $\sum_k \hat{\beta}_{k,i} \bar{f}_k$, where $\bar{f}_k$ is the original-sample mean return of factor k. Fama and French (1993) state that $\hat{\beta}_{i,k}$ with respect to their SMB and HML factors have "a clear interpretation as risk-factor sensitivities." If this interpretation is both correct and stable, then the estimated expected return should remain post-sample.

Figure 8 plots the post-sample mean return against the factor model expected returns. We normalize by the original-sample mean return for ease of interpretation. With this normalization, the position on the x-axis ([Predicted by Risk Model]/[In-Sample]) represents the share of predictability due to risk.

The figure shows that a minority of in-sample predictability is attributed to risk, at best. Using the CAPM (Panel (a)), nearly all predictability is less than 25% due to risk

Figure 8: Mean Returns Post-Sample vs Factor Model Predictions

Each marker is one published long-short strategy. [Post-Sample]/[In-Sample] is the mean return post-sample divided by the mean return in-sample. [Predicted by Risk Model] is $\sum_k \hat{\beta}_{k,i} \bar{f}_k$, where $\bar{f}_k$ is the in-sample mean return of factor k and $\hat{\beta}_{k,i}$ comes from an in-sample time series regression of long-short returns on factor realizations. FF3 and FF5 are the Fama-French 3- and 5-factor models. The blue line is the OLS fit. The axes zoom in on the interpretable region of the chart and omits outliers. Factor models attribute a minority of in-sample predictability to risk, at best. Post-sample decay is the distance between the horizontal line at 1.0 and the regression line, and this decay is near 50% even for predictors that are entirely due to risk according to the CAPM and FF3. For FF5, decay is smaller for predictors that are more than 75% due to risk, but these predictors are rare.

(a) Post-Sample Return vs CAPM Prediction   (b) Post-Sample Return vs FF3 Prediction

(c) Post-Sample Return vs FF5 Prediction

(to the left of the vertical line at 0.25), and many predictors have a *negative* risk share. FF3 (Panel (b)) implies more predictability is due to risk, but still the vast majority of predictors lie to the left of 0.50.

Fama and French (2015) are more cautious than Fama and French (1993), and describe the risk-based ICAPM as "the more ambitious interpretation" of the five factor model. Under the more ambitious interpretation, FF5 implies that most predictors are less than 50% due to risk. These results are consistent with our manual reading of the papers, which typically attribute predictability to mispricing (Table 4).

The regression lines in Figure 8 show negative or mildly positive relationships between factor model risk and post-sample returns. The regression fits for the CAPM and FF3 models never stray far from 50%, implying that even predictors that are entirely due to risk are little different than the typical predictor in terms of post-sample robustness. FF5 risk shows a stronger relationship with post-sample returns, but even the rare predictors that are 75% due to risk decay by roughly 40% post-sample. Moreover, the Fama and French (2015) model may have the benefit of hindsight, as the median publication year for the Chen-Zimmermann predictors is 2006.

6   Conclusion

The peer review process demands a significant investment of talent and energy. This paper examines whether this investment helps predict the cross-section of stock returns. We find that, compared to naively searching for statistically significant accounting ratios, the additional predictive power is modest. Post-sample returns from the peer review process and data mining are quite similar, with peer review out-performing by about 2 bps per month. Additionally, data mining mimics features of the peer review process, including subtle patterns in event time returns and themes like investment and issuance. These results suggest that the process for generating peer-reviewed predictors is, itself, data mining.

Research that focuses on risk-based explanations or equilibrium models also fail to out-perform. If anything, these arguably more rigorous methods lead to worse performance. We provide a meta theory for this phenomenon, which shows that using risk-based explanations either fails to provide a signal of higher expected returns or fails to provide a signal of stable expected returns, compared to other methods like behavioral explanations or data mining. This result is consistent with surveys of real-world investors, who consistently overlook academic risk factors in their investing decisions. More positively, we find that peer review attributes only 18% of cross-sectional predictors to risk.

Our findings are also positive for the growing literature on machine learning in finance. If naive data mining can generate substantial out-of-sample returns, then more sophisticated machine learning methods should have even more potential. This result

suggests that a way forward in asset pricing is to let the data speak directly, without the filters from traditional theory, following the lead of fields like protein folding and linguistics.

Appendix A

A.1 Risk words and mispricing words

We remove stopwords, lowercase and lemmatize all words using standard methods. Then, we count separately the words corresponding to risk and mispricing.

We consider as risk words the following terms and their grammatical variations: "utility," "maximize," "minimize," "optimize," "premium," "premia," "premiums," "consume," "marginal," "equilibrium," "sdf," "investment-based," and "theoretical." We also count as risk words appearances of "risk" that are not preceded by "lower," and appearances of "aversion," "rational," and "risky" that are not preceded by "not."

The mispricing words consist of " "anomaly," "behavioral," "optimistic," "pessimistic," "sentiment," "underreact," "overreact," "failure," "bias," "overvalue," "misvalue," "undervalue," "attention," "underperformance," "extrapolate," "underestimate," "misreaction," "inefficiency," "delay," "suboptimal," "mislead," "overoptimism," "arbitrage," "factor unlikely," and their grammatical variations. We further count as mispricing the terms "not rewarded," "little risk," "risk cannot [explain]," " low [type of] risk," "unrelated [to the type of] risk," "fail [to] reflect," and "market failure," where the terms in brackets are captured using regular expressions or correspond to stopwords.

A.2 Additional Empirical Results

Figure A.1: Agnostic and Mispricing Predictors vs Data-Mining

The plot shows long-short returns in event time, where the event is the end of the original sample periods. Predictor returns are normalized to average 100 bps in the original samples. Data-mined predictors come from ratios or scaled first differences from 240 accounting variables (Section 2.1.1). For all categories of theory, peer review and data mining lead to similar post-sample returns.

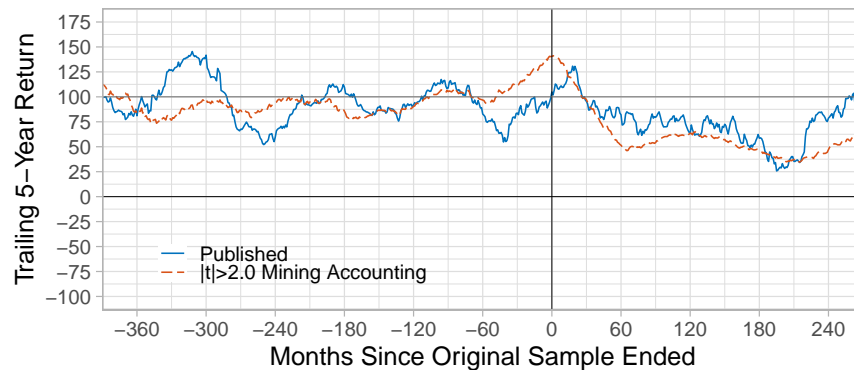(a) Mispricing-Based
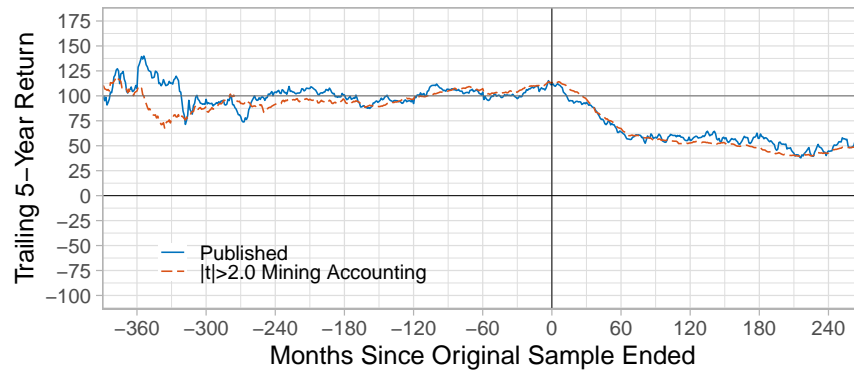
(b) Agnostic

Table A.1: 20 Data-Mined Predictors That Perform Similarly to Banz's Size (1981)

Table lists 20 of the 221 data-mined signals that performed similarly to Banz's (1981) size in the original sample period. Signals are ranked according to the absolute difference in mean original-sample return. Sign = -1 indicates that a high signal implies a lower mean return original-sample. Data mining leads to similar out-of-sample performance.

| Similarity Rank	Signal	Sign	Mean Return (% Monthly) 1926-1975	1976-2021
Peer-Reviewed				
	Size (Banz 1981)	-1	0.50	0.19
Data-Mined				
1	Δ[Equity liq value]/lag[Sales]	-1	0.50	0.77
2	Δ[Invest tax credit inc ac]/lag[Pref stock redemp val]	-1	0.49	-0.13
3	[Cost of goods sold]/[Capex PPE sch V]	1	0.50	0.82
4	Δ[Assets]/lag[Preferred stock liquidat]	-1	0.49	0.20
5	Δ[Equity liq value]/lag[Curr assets other sundry]	-1	0.48	0.73
6	Δ[Equity liq value]/lag[Current liabilities]	-1	0.48	0.85
7	Δ[Receivables]/lag[Pref stock redemp val]	-1	0.48	0.19
8	[Market equity FYE]/[Invested capital]	-1	0.49	0.73
9	Δ[Current assets]/lag[Invest tax credit inc ac]	-1	0.52	0.33
10	Δ[Assets]/lag[Pref stock redemp val]	-1	0.47	0.25
	. . .			
101	Δ[Sales]/lag[Current liabilities]	-1	0.40	0.77
102	Δ[Interest expense]/lag[Cost of goods sold]	-1	0.40	0.71
103	Δ[Interest expense]/lag[Num employees]	-1	0.40	0.68
104	Δ[Operating expenses]/lag[Long-term debt]	-1	0.40	0.48
105	Δ[Operating expenses]/lag[Current liabilities]	-1	0.40	0.90
	. . .			
234	[Gross profit]/[Earnings before interest]	1	0.35	0.19
235	[Market equity FYE]/[Capex PPE sch V]	-1	0.35	0.30
236	[PPE land and improvement]/[Pension retirem expense]	-1	0.64	-0.00
237	[Interest expense]/[Cost of goods sold]	-1	0.35	0.63
238	[Operating expenses]/[Op income after deprec]	1	0.35	0.15
	Mean Data-Mined		0.43	0.44

Table A.2: Signals by Theory and Published Journal

This table lists the number of signals by theory and published journal. Finance journals find risk explanations more frequently than accounting journals, but risk explanations still account for a small minority of predictors in finance journals.

	Agnostic	Mispricing	Risk
AR	1	14	0
BAR	0	1	0
Book	2	0	0
CAR	0	1	0
FAJ	1	1	0
JAE	3	14	0
JAR	3	2	0
JBFA	0	1	0
JEmpFin	0	1	0
JF	16	35	12
JFE	16	22	6
JFM	0	2	0
JFQA	0	3	2
JFR	0	0	1
JOIM	0	1	0
JPE	0	0	3
JPM	1	0	0
MS	0	2	2
Other	0	1	0
RAS	0	5	1
RED	0	0	1
RFQA	0	1	0
RFS	0	7	7
ROF	0	1	3
WP	1	1	0

Figure A.2: Abnormal CAPM Returns

The plot shows the abnormal return of long-short returns of published predictors in event time, where the event is the end of the original sample periods. We calculate abnormal returns as $abnormal_{i,t} = r_{i,t} - \beta_{i,t} r_t^e$. We calculate beta separately for the original sample period, and after the original sample period. We keep the abnormal returns if the t-statistic is greater than one during the original sample period. Each abnormal return is normalized so that its mean original-sample is 100 bps per month. Predictors are grouped by theory category based on the arguments in the original papers (Table 3). We average abnormal returns across predictors within each month and then take the trailing 5-year average for readability. For all categories of theory, predictability decays by roughly 50% post-sample. If anything, risk-based predictors decay more than other predictors.
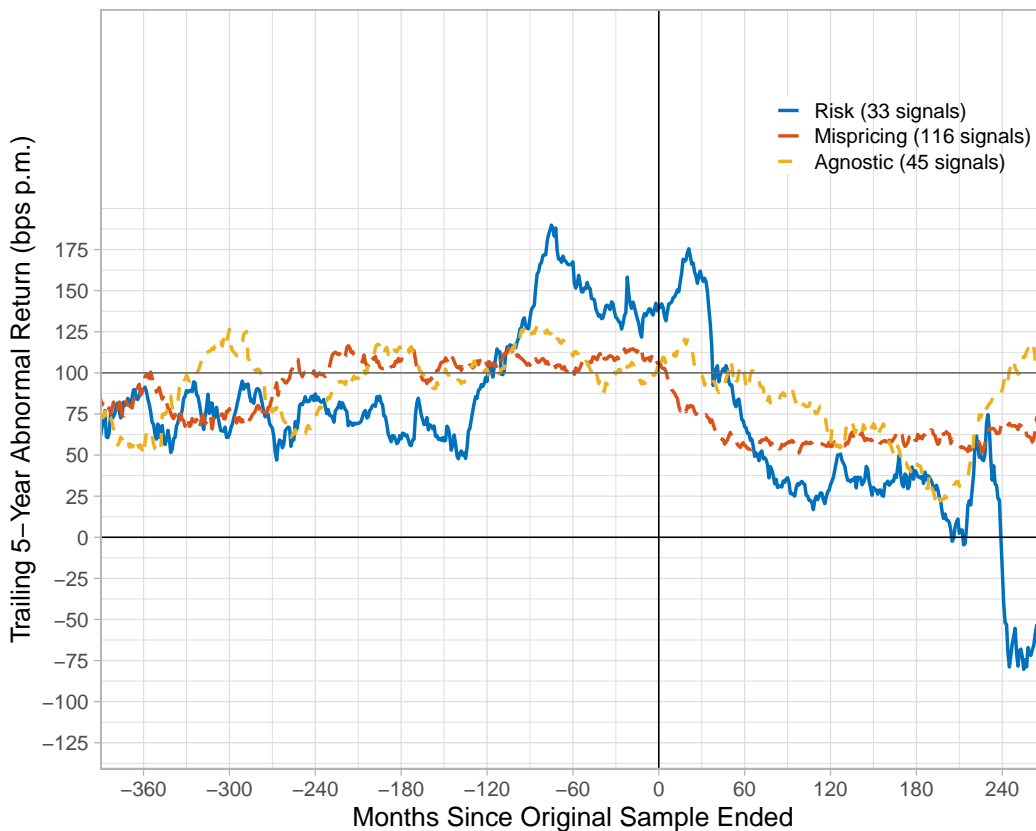
A.3 Proof of Proposition 1

Proof. Let $s_i \equiv \text{Sign}(\bar{r}_i)$. Then rewrite the expected decay conditioning on $|\bar{r}_i|$ and i belonging to some set $\mathcal{S}$

$$E\left[\text{Decay}_i \mid |\bar{r}_i|, i \in \mathcal{S}\right] = E\left[s_i\left(\bar{r}_i - \bar{r}_i^{OOS}\right) ||\bar{r}_i|, i \in \mathcal{S}\right] \tag{7}$$

$$= E\left[s_i\left(\bar{r}_i - \mu_i - \Delta\mu_i - \bar{\varepsilon}_i^{OOS}\right) ||\bar{r}_i|, i \in \mathcal{S}\right] \tag{8}$$

$$= E\left[s_i\left(\bar{r}_i - \mu_i - \Delta\mu_i\right) ||\bar{r}_i|, i \in \mathcal{S}\right] \tag{9}$$

$$= E\left[s_i\bar{r}_i - s_i\mu_i - s_i\Delta\mu_i ||\bar{r}_i|, i \in \mathcal{S}\right] \tag{10}$$

where the first line uses Equation (1) and the second line uses Equation (3).

Then plug in $\mathcal{S} = \mathcal{T}$ and then use the two assumptions of the proposition (Equations (5) and (6))

$$E\left[\text{Decay}_i \mid |\bar{r}_i|, i \in \mathcal{T}\right] = E\left[s_i\bar{r}_i - s_i\mu_i + s_i\Delta\mu_i \mid \bar{r}_i, i \in \mathcal{T}\right]$$

$$< E\left[s_i\bar{r}_i - s_i\mu_i + s_i\Delta\mu_i \mid \bar{r}_i, i \in \mathcal{D}\right]$$

$$= E\left[\text{Decay}_i \mid |\bar{r}_i|, i \in \mathcal{D}\right]$$

Integrating over $|\bar{r}_i| > h$ finishes the proof. $\square$

References

Abarbanell, Jeffery S and Brian J Bushee (1998). "Abnormal returns to a fundamental analysis strategy". In: *Accounting Review*, pp. 19–45.

Amihud, Yakov (2002). "Illiquidity and stock returns: cross-section and time-series effects". In: *Journal of financial markets* 5.1, pp. 31–56.

Ankel-Peters, Jörg, Nathan Fiala, and Florian Neubauer (2023). "Is Economics Self-Correcting? Replications in the American Economic Review". In.

Bali, Turan G, Heiner Beckmeyer, and Timo Wiedemann (2023). "Expected Mispricing". In: *Available at SSRN*.

Bali, Turan G, Robert F Engle, and Scott Murray (2016). *Empirical asset pricing: The cross section of stock returns*. John Wiley & Sons.

Banz, Rolf W (1981). "The relationship between return and market value of common stocks". In: *Journal of financial economics* 9.1, pp. 3–18.

Barberis, Nicholas (2018). "Psychology-based models of asset prices and trading volume". In: *Handbook of behavioral economics: applications and foundations 1*. Vol. 1. Elsevier, pp. 79–175.

Barry, Christopher B and Stephen J Brown (1984). "Differential information and the small firm effect". In: *Journal of financial economics* 13.2, pp. 283–294.

Bender, Svetlana et al. (2022). "Millionaires speak: What drives their personal investment decisions?" In: *Journal of Financial Economics* 146.1, pp. 305–330.

Bessembinder, Hendrik, Aaron Burt, and Christopher M Hrdlicka (2021). "Time Series Variation in the Factor Zoo". In: *Aaron Paul and Hrdlicka, Christopher M., Time Series Variation in the Factor Zoo (December 22, 2021)*.

Böll, Julian et al. (2024). "Following the Footprints: Towards a Taxonomy of the Factor Zoo". In: *Available at SSRN*.

Boudoukh, Jacob et al. (2007). "On the importance of measuring payout yield: Implications for empirical asset pricing". In: *The Journal of Finance* 62.2, pp. 877–915.

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

Celerier, Claire, Boris Vallee, and Alexey Vasilenko (2022). "What Drives Finance professors' Wages?" In.

Chen, Andrew Y (2018). "A general equilibrium model of the value premium with time-varying risk premia". In: *The Review of Asset Pricing Studies* 8.2, pp. 337–374.

— (2021). "The Limits of p-Hacking: Some Thought Experiments". In: *The Journal of Finance* 76.5, pp. 2447–2480.

Chen, Andrew Y (2024). "Most claimed statistical findings in cross-sectional return predictability are likely true". In: *arXiv preprint arXiv:2206.15365*.

Chen, Andrew Y and Chukwuma Dim (2023). "High-Throughput Asset Pricing". In: *arXiv preprint arXiv:2311.10685*.

Chen, Andrew Y and Jack McCoy (2024). "Missing values handling for machine learning portfolios". In: *Journal of Financial Economics* 155, p. 103815.

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.

— (2023). "Publication Bias in Asset Pricing Research". In: *Oxford Research Encyclopedia of Economics and Finance*.

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

Chinco, Alex, Samuel M Hartzmark, and Abigail B Sussman (2022). "A new test of risk factor relevance". In: *The Journal of Finance* 77.4, pp. 2183–2238.

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 (2009). *Asset pricing: Revised edition*. Princeton university press.

— (2017). "Macro-finance". In: *Review of Finance* 21.3, pp. 945–985.

Cooper, Michael J, Huseyin Gulen, and Michael J Schill (2008). "Asset growth and the cross-section of stock returns". In: *the Journal of Finance* 63.4, pp. 1609–1651.

DeMiguel, Victor et al. (2020). "A Transaction-Cost Perspective on the Multitude of Firm Characteristics". In: *The Review of Financial Studies* 33.5, pp. 2180–2222.

Doran, James and Colbrin Wright (2007). "What Really Matters When Buying and Selling Stocks?" In: *Financial Education* 8.1, pp. 35–61.

Fama, Eugene F (1970). "Efficient capital markets: A review of theory and empirical work". In: *The journal of Finance* 25.2, pp. 383–417.

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.

— (2006). "Profitability, investment and average returns". In: *Journal of financial economics* 82.3, pp. 491–518.

Fama, Eugene F and Kenneth R French (2015). "A five-factor asset pricing model". In: *Journal of financial economics* 116.1, pp. 1–22.

— (2018). "Choosing factors". In: *Journal of financial economics* 128.2, pp. 234–252.

Frazzini, Andrea and Lasse Heje Pedersen (2014). "Betting against beta". In: *Journal of Financial Economics* 111.1, pp. 1–25.

Frey, Jonas (2023). "Which stock return predictors reflect mispricing?" In: *Available at SSRN*.

Freyberger, Joachim, Andreas Neuhierl, and Michael Weber (2020). "Dissecting characteristics nonparametrically". In: *The Review of Financial Studies* 33.5, pp. 2326–2377.

Goto, Shingo and Toru Yamada (2022). "False Alpha and Missed Alpha: An Out-of-Sample Mining Expedition". In: *Working Paper*.

Green, Jeremiah, John RM Hand, and X Frank Zhang (2017). "The characteristics that provide independent information about average US monthly stock returns". In: *The Review of Financial Studies* 30.12, pp. 4389–4436.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu (2020). "Empirical asset pricing via machine learning". In: *The Review of Financial Studies* 33.5, pp. 2223–2273.

Han, Yufeng et al. (2022). "Expected Stock Returns and Firm Characteristics: E-ENet, Assessment, and Implications". In: *Working Paper*.

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

Haugen, Robert A and Nardin L Baker (1996). "Commonality in the determinants of expected stock returns". In: *Journal of financial economics* 41.3, pp. 401–439.

Heston, Steven L and Ronnie Sadka (2008). "Seasonality in the cross-section of stock returns". In: *Journal of Financial Economics* 87.2, pp. 418–445.

Holcblat, Benjamin, Abraham Lioui, and Michael Weber (2022). "Anomaly or possible risk factor? Simple-to-use tests". In: *Simple-To-Use Tests (April 3, 2022)*.

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, Michael C. and George A. Benington (1970). "Random Walks and Technical Theories: Some Additional Evidence". In: *The Journal of Finance* 25.2, pp. 469–482.

Jensen, Theis Ingerslev, Bryan T Kelly, et al. (2022). "Machine learning and the implementable efficient frontier". In: *Swiss Finance Institute Research Paper* 22-63.

Jumper, John et al. (2021). "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873, pp. 583–589.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh (2020). "Shrinking the cross-section". In: *Journal of Financial Economics* 135.2, pp. 271–292.

Krusell, Per and Anthony A Smith Jr (1998). "Income and wealth heterogeneity in the macroeconomy". In: *Journal of political Economy* 106.5, pp. 867–896.

Lo, Andrew W and A Craig MacKinlay (1990). "Data-snooping biases in tests of financial asset pricing models". In: *The Review of Financial Studies* 3.3, pp. 431–467.

Lopez-Lira, Alejandro and Nikolai L Roussanov (2020). "Do Common Factors Really Explain the Cross-Section of Stock Returns?" In: *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper*.

Marcet, Albert (1991). "Solving non-linear stochastic models by parameterizing expectations: An application to asset pricing with production". In.

Mas-Colell, Andreu, Michael Dennis Whinston, and Jerry R Green (1995). *Microeconomic theory*. Vol. 1. Oxford university press New York.

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

Moritz, Benjamin and Tom Zimmermann (2016). "Tree-based conditional portfolio sorts: The relation between past and future stock returns". In: *Available at SSRN 2740751*.

Moskowitz, Tobias J and Mark Grinblatt (1999). "Do industries explain momentum?" In: *The Journal of finance* 54.4, pp. 1249–1290.

Mukhlynina, Liliya and Kjell G Nyborg (2020). "The Choice of Valuation Techniques in Practice: Education Versus Profession". In: *Critical Finance Review* 9.1-2, pp. 201–265.

Pástor, Ľuboš and Robert F Stambaugh (2003). "Liquidity risk and expected stock returns". In: *Journal of Political economy* 111.3, pp. 642–685.

Pontiff, Jeffrey and Artemiza Woodgate (2008). "Share issuance and cross-sectional returns". In: *The Journal of Finance* 63.2, pp. 921–945.

Shiller, Robert J (2003). "From efficient markets theory to behavioral finance". In: *Journal of economic perspectives* 17.1, pp. 83–104.

Sloan, Richard G (1996). "Do stock prices fully reflect information in accruals and cash flows about future earnings?" In: *Accounting review*, pp. 289–315.

Soliman, Mark T (2008). "The use of DuPont analysis by market participants". In: *The Accounting Review* 83.3, pp. 823–853.

Stattman, Dennis (1980). "Book values and stock returns". In: *The Chicago MBA: A journal of selected papers* 4.1, pp. 25–45.

Sullivan, Ryan, Allan Timmermann, and Halbert White (1999). "Data-snooping, technical trading rule performance, and the bootstrap". In: *The journal of Finance* 54.5, pp. 1647–1691.

— (2001). "Dangers of data mining: The case of calendar effects in stock returns". In: *Journal of Econometrics* 105.1, pp. 249–286.

Tuzel, Selale (2010). "Corporate real estate holdings and the cross-section of stock returns". In: *The Review of Financial Studies* 23.6, pp. 2268–2302.

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.

Zaffaroni, Paolo and Guofu Zhou (2022). "Asset Pricing: Cross-section Predictability". In: *Available at SSRN 4111428*.

Zhang, Lu (2005). "The value premium". In: *The Journal of Finance* 60.1, pp. 67–103.

Zhao, Wayne Xin et al. (2023). *A Survey of Large Language Models*. arXiv: `2303.18223 [cs.CL]`.

Zhu, Min (2023). "Evaluating the Efficacy of Multiple Testing Adjustments in Empirical Asset Pricing". In: *Available at SSRN 4396035*.

CFR Working Paper Series

centre for financial research cologne

CFR working papers are available for download from www.cfr-cologne.de.

2024

No.	Author(s)	Title
24-02	A. Y. Chen, A. Lopez-Lira, T. Zimmermann	Does Peer-Reviewed Research Help Predict Stock Returns?
24-01	G. Cici, P. Schuster, F. Weishaupt	Once a Trader, Always a Trader: The Role of Traders in Fund Management

2023

No.	Author(s)	Title
23-08	A. Braun, J. Braun, F. Weigert	Extreme Weather Risk and the Cost of Equity
23-07	A. G. Huang, R. Wermers, J. Xue	"Buy the Rumor, Sell the News": Liquidity Provision by Bond Funds Following Corporate News Events
23-06	J. Dörries, O. Korn, G. J. Power	How Should the Long-term Investor Harvest Variance Risk Premiums?
23-05	V. Agarwal, W. Jiang, Y. Luo, H. Zou	The Real Effect of Sociopolitical Racial Animus: Mutual Fund Manager Performance During the AAPI Hate
23-04	V. Agarwal, B. Barber, S. Cheng, A. Hameed, H. Shanker, A. Yasuda	Do Investors Overvalue Startups? Evidence from the Junior Stakes of Mutual Funds
23-03	A. Höck, T. Bauckloh, M. Dumrose, C. Klein	ESG Criteria and the Credit Risk of Corporate Bond Portfolios
23-02	T. Bauckloh, J. Dobrick, A. Höck, S. Utz, M. Wagner	In partnership for the goals? The level of agreement between SDG ratings
23-01	F. Simon, S. Weibels, T. Zimmermann	Deep Parametric Portfolio Policies

2022

No.	Author(s)	Title
22-12	V. Agarwal, A. Cochardt, V. Orlov	Birth Order and Fund Manager's Trading Behavior: Role of Sibling Rivalry
22-11	G. Cici, S. Gibson, N. Qin, A. Zhang	The Performance of Corporate Bond Mutual Funds and the Allocation of Underpriced New Issues

This document only covers the most recent CFR working papers. A full list can be found at www.cfr-cologne.de.

centre for financial research
cfr/university of cologne
Albertus-Magnus-Platz
D-50923 cologne
fon +49(0)221-470-6995
fax +49(0)221-470-3992
kempf@cfr-cologne.de
www.cfr-cologne.de