Highlights

- We stress-test MTPs with the aim of finding their breaking point under extreme data snooping efforts.
- MTP size distortions depend on methodological choices and data sample properties.
- MTPs fail to provide finite-sample control of aggregate error rates in bearish-leaning, less volatile samples, irrespective of sample length.
- On average, a t-ratio of 3.85 is required to control for lucky trading rules in bearish samples, representing a two-fold increase compared to bullish samples; this may increase with poor methodological choices.
- When controlling for asymmetric data snooping bias, speculative trading rules earn next to no economic profits for traders in the cryptocurrency, stock, and foreign exchange markets.

When Multiple Testing Procedures Fail Under Extreme Data Snooping Pressure*

Dan Gabriel Anghel[a,b,*]

[a] *Institute for Economic Forecasting, Romanian Academy*
[b] *Faculty of Finance and Banking, The Bucharest University of Economic Studies*

Abstract

We perform a large scale stress test of modern Multiple Testing Procedures (MTPs) that are used to evaluate the superior predictive ability of many forecasting models, documenting how their size depends on methodological choices and data sample properties in environments plagued by data snooping. We specifically focus on the evaluation of technical trading rules, whose number has been exponentially increasing in recent years. We find that false discoveries (Type I errors) increase when the sample average return and volatility decrease. Notably, MTPs fail to provide adequate finite-sample control of aggregate error rates when tests are performed on bearish data samples or when market frictions (trading fees, liquidity costs, short selling restrictions) are ignored. To control for asymmetric data snooping bias, researchers should increase the statistical significance threshold used to detect superior forecasting performance in downward trending markets.

*This work was supported by a grant of the Romanian Ministry of Education and Research, CNCS - UEFISCDI, project number PN-III-Pl-l.1-PD-2019-0016, within PNCDI III.

*13 Calea 13 Septembrie, Bucharest, Romania.
Email address: dan.anghel@fin.ase.ro (Dan Gabriel Anghel)

JEL Classification: C12, C18, G11, G14

Keywords: Multiple Testing Procedures, Forecasting Models, Technical Trading Rules, Superior Predictive Ability, Data Snooping Bias, False Discoveries

---

1. Introduction

One of the common aspirations of academic researchers, investment professionals, and individual traders is understanding and predicting financial asset price movements. As computational resources have become increasingly accessible, the number of models that claim superior predictive ability and the associated number of trading rules that claim superior economic profitability has considerably increased. In this context, it is crucial for stakeholders to have at their disposal and to use adequate statistical tools, Multiple Testing Procedures (MTPs) that weed out false discoveries and help eliminate models who's in-sample performance is simply due to luck from truly economically superior models. Equally important is to understand the empirical limitations of existing MTPs, given the current environment characterized by extreme data snooping pressure. However, despite the long-running discussion on lucky trading performance (see, e.g., the seminal work of Cowles 3rd, 1933) and data snooping (e.g. Brock, Lakonishok, and LeBaron, 1992), many details regarding how and when false discoveries arise or regarding their prevalence in empirical tests remain unclear. For example, lucky trading performance is typically associated with the spurious ability of models to make forecasts that correlate with future asset returns by chance, but the actual correspondence between this property and the amount of false discoveries (Type I errors) has not been specifically evaluated so far. Could test results be influenced by other trading rule characteristics? Also, the effect of ignoring trading costs on trading rule overperformance has been extensively discussed (Fama, 1965, provides an early example) but its influence of the amount of false discoveries has not been precisely quantified, especially

in the context of modern MTPs that claim to control for data snooping. More generally, there is close to no evidence on the resilience of MTPs to extreme data snooping pressure under different testing conditions[1], and especially to intentional or unintentional test misspecification.

In this paper, we fill this gap by extensively investigating the drivers of false discoveries in empirical tests that use MTPs to evaluate the performance of forecasting models and associated trading rules in financial markets, motivated by the observation that this environment is plagued by data snooping efforts. Specifically, we use a very large set of technical trading rules (TTRs) to investigate the factors (data sample properties, methodological choices, and rule characteristics) that bias MTP results, and estimate the relative contribution of each factor to the aggregate bias. In essence, we conduct a practically-inspired large-scale stress test of existing MTPs, which aims to document how and when they might fail, and what are the remedies that researchers can apply to assure robust inferences.

Overall, we show that seemingly minor changes that empirical researchers explicitly or implicitly make when applying tests in practice provoke significant differences in results and could lead to false discoveries. The main novel finding is that a *spurious bearish tendency* of lucky trading rules in samples exhibiting negative average returns is the main driver of false discoveries, while *spurious correlation* is less important. In samples with strongly nega-

[1]We use *testing conditions* in reference to the combination of data sample properties (sample length, sample average return, standard deviation of returns, etc.) and methodological choices that researchers make (accounting for trading costs and/or short trading restrictions, standardizing the test statistic, handling for "deep-in-the-null" trading rules as in Hansen (2005) or Romano and Wolf (2018), etc.).

tive average returns, false discoveries also increase when volatility decreases. More generally, we show that size distortions of MTPs under extreme data snooping pressure depend on the properties of sample returns. This holds for a wide variety of methodological choices, and can be observed in both Monte Carlo simulations and empirical tests conducted on the cryptocurrency, stock, and foreign exchange markets, respectively. We further find that false discoveries due to spurious correlation can be eliminated when accounting for all data snooping efforts (i.e., when considering the entire trading rule set from which lucky rules are extracted), while false discoveries due to spurious bearish tendency are much more persistent and cannot be eliminated using existing methodological approaches when relying on standard significance thresholds.

On the one hand, the results imply that the sources of data snooping bias are more diverse than previously recognized and that methodological precautions are required to adequately control for lucky trading performance when specific testing conditions are met. For example, when applying MTPs on samples with bearish characteristics, a t-ratio of 3.85 is needed on average to provide finite-sample control for aggregate error rates at the 5% level (3.21 for the 10% level and 5.33 for the 1% level), representing a two-fold increase compared to samples with bullish characteristics. On the other hand, the key implication for the literature examining TTR overperformance is that previous positive discoveries obtained on samples with negative average returns may be false and should be reevaluated after controlling for asymmetric data snooping bias. This would include short samples containing market crashes or, more generally, bearish price movements, but also longer samples col-

lected for financial assets that do not have a theoretical positive price drift, such as derivatives, currencies, or cryptocurrencies. For example, a growing body of evidence based on modern MTPs has shown that data snooping bias is largely responsible for positive discoveries regarding TTR overperformance in the stock market (see Anghel, 2021a, and references therein). However, this is not the case in the foreign exchange or cryptocurrency markets, where various recent contributions still support the superior forecasting ability of TTRs (e.g., Zarrabi, Snaith, and Coakley, 2017; Corbet, Eraslan, Lucey, and Sensoy, 2019; Grobys, Ahmed, and Sapkota, 2020). For the latter, researchers also typically document a pattern of time-varying TTR performance, casting doubt on the efficiency of short-term price movements in specific time periods. Even though not directly challenging previous findings, our analysis hints that asymmetric data snooping bias may play a role in these and related studies, and that the results should be reevaluated after explicitly taking it into account.

Our paper is mainly related to the replication crisis in finance and the discussion on lucky forecasting models that has recently gained significant traction in the literature (see, e.g., the presidential address of Harvey, 2017). Although the discussion is not new, evidence showing how much data snooping bias we could expect in empirical tests only recently emerged. For example, Dichtl, Drobetz, Neuhierl, and Wendt (2021) find that almost all strategies designed to forecast equity risk premiums fail to beat the simple historical mean in out-of-sample tests after controlling for data snooping and accounting for transaction costs. Similarly, Harvey, Liu, and Saretto (2020) show that the vast majority of trading strategies based on fundamentals

have no economically-significant forecasting ability for the cross section of stock returns. McLean and Pontiff (2016), Linnainmaa and Roberts (2018) or Chordia, Goyal, and Saretto (2020) go one step further and estimate that the proportion of false discoveries published in the asset pricing literature may range between 26% and 58%, implying that data snooping severely biases our perception about equity risk factors and fund managers skill. More generally, Harvey, Liu, and Zhu (2016) argue that most positive findings in financial economics are likely false and suggest that a newly discovered, non-theory-based factor needs to clear a much higher statistical hurdle, i.e. a t-ratio ≥ 3.

Regarding technical (speculative) trading rules, i.e. active trading strategies that are based solely on historical asset prices, Anghel (2021a) estimates that between 50% and 75% of results supporting superior forecasting (trading) performance may be false, as researchers usually ignore their combined data snooping efforts. However, Anghel (2021a) also finds that false discoveries persist even after accounting for all data snooping, hinting that other factors may also be involved, but does not investigate further. To the extent of our knowledge, no paper focusing on the performance of trading rules has analyzed why/when is data snooping bias persistent, or performed an in-depth investigation into the exact contribution of each possible explanatory factor to the total bias. Here, we perform such an investigation by analyzing the testing conditions that make trading rules appear more profitable than they truly are.

The remainder of the paper is organized as follows. Section 2 discusses the statistical properties of excess performance measures and possible sources

of data snooping bias. Section 3 reports and discusses the stress test results, i.e. the results of several Monte Carlo simulations that analyze the factors that drive false discoveries in MTPs. It also includes robustness checks that confirm the *spurious bearish tendency* effect as the main factor that biases results. Section 4 shows that spurious bearish tendency can also be observed in empirical tests conducted on the cryptocurrency, stock, and foreign exchange markets, respectively, regardless of the size of the data sample used. The analysis also reveals that speculative trading rules have next to no economically significant superior predictive ability after taking asymmetric data snooping bias into account. Section 5 concludes.

2. Trading rule overperformance, statistical tests, and potential sources of data snooping bias

Various testing procedures can be used to evaluate the economic performance of speculative trading rules. However, because many rules are explicitly or implicitly evaluated together (Sullivan, Timmermann, and White, 1999), researchers should account for the possibility of lucky overperformance by handling for the associated multiple hypotheses. Approaches that do this include the False Discovery Rate (FDR) test (Benjamini and Hochberg, 1995) and its extensions (e.g. Storey, 2002; Bajgrowicz and Scaillet, 2012), or the Reality Check test (White, 2000) and its extensions (e.g. Hansen, 2005; Romano and Wolf, 2005; Hsu, Hsu, and Kuan, 2010; Hansen, Lunde, and Nason, 2011). Harvey et al. (2020) provide a comprehensive review of relevant MTPs that are useful for finance applications.

Are modern testing procedures immune to data snooping bias? Surely

not, as the way and context in which they are applied matters. Indeed, Anghel (2021a) shows that false discoveries arise when researchers misspecify the set of alternative trading rules against which they check for statistically significant excess performance. Moreover, false discoveries persist even after accounting for all data snooping efforts (Anghel, 2021a). Why is data snooping bias so persistent, and what are the conditions in which we can expect it to rise up to non-negligible levels? Similarly, it is well known that classical test results can be biased when ignoring trading costs and other market frictions, the literature on this going back at least to Fama (1965). However, there is no evidence showing exactly how ignoring market frictions impacts MTP results. To provide additional insights, we first analyze the way in which the economic performance of speculative trading rules is evaluated.

At their core, MTPs evaluate the null hypothesis of no overperformance, which can be stated as $H_0 : \max_{k=1..K} \mathbb{E}[d_k] \leq 0$, where d_k is the excess performance associated with the gain (negative loss) function of a trading rule or, more generally, of a forecasting model k, and K is the number of models being considered simultaneously. In the context of speculative rules used by traders in financial markets, the gain function is typically linear, being defined as the excess return over the buy-and-hold strategy ($\delta_{0,t} = 1, \forall t$), which is used as a benchmark because it represents the optimal alternative when prices are not predictable, e.g. follow a random walk. Thus, and accounting for transaction costs, the gain function can be defined as:

$$d_{k,t} = [\delta_{k,t-1}\xi_t - \mathbb{1}_{\{\delta_{k,t-1} \neq \delta_{k,t-2}\}}(\phi + \lambda_t)] - \delta_{0,t-1}\xi_t \tag{1}$$

where ξ_t denotes the market return, ϕ denotes the fixed broker fee, λ_t denotes

the liquidity and/or price impact cost, and δ_k is the signal function associated with trading rule k.[2] The term in square brackets represents the absolute performance (return), while subtracting the market return yields the relative, excess performance (return). Taking expectations results in the following decomposition of trading rule expected excess performance:

$$\mathbb{E}[d_k] = \rho_{\delta,\xi}\sigma_\delta\sigma_\xi + (\mathbb{E}[\delta_k] - 1)\mathbb{E}[\xi] - \varphi \tag{2}$$

where $\sigma_\delta = \sqrt{Var(\delta_k)} > 0$ denotes the volatility of trading signals, $\sigma_\xi = \sqrt{Var(\xi)} > 0$ denotes the volatility of market returns, $\rho_{\delta,\xi} = Corr(\delta_k, \xi) \in [0, 1]$ denotes the correlation between (lagged) trading signals[3] and market returns, i.e., the rule's *predictive accuracy*, $\varphi = \theta(\phi + \mathbb{E}[\lambda]) + \sigma_{\delta,\lambda} > 0$, denotes the total trading cost penalty, with $\theta = \mathbb{P}(\delta_{k,t-1} \neq \delta_{k,t-2}) \in [0, 1]$ being the probability of making a trade, i.e. the rule's *trading frequency*, and $\sigma_{\delta,\lambda} = Cov(\mathbb{1}_{\{\delta_{k,t-1} \neq \delta_{k,t-2}\}}, \lambda_t) > 0$ being the covariance between trade occurrences and liquidity costs, i.e. the rule's *illiquidity preference*.

When trading in a market that explicitly disallows short positions or in which short trades are effectively not functional, only long trades are possible and the signal function can take either 1 (with probability $p_k \in [0, 1]$) or

[2]Signal functions are step functions that use historical prices (or returns) to predict the future direction of price movements and instruct traders about what positions they should take in the market, i.e. -1 when prices are expected to decrease and traders should go short, 0 when prices are expected to remain constant and traders should stay out of the market, or 1 when prices are expected to increase and traders should go long. In practice, unchanged prices are hardly predicted, meaning we can discard this possibility from the analysis and model signals as *Bernoulli* random variables without a loss of generality. However, as discussed below, 0 can still appear as a signal when prices are expected to decrease but short trades are explicitly or implicitly not allowed.

[3]Lagged signals are used to avoid hindsight bias;

0 (with probability $1 - p_k \in [0, 1]$). In this case, δ_k is a *Bernoulli* random variable with mean $\mathbb{E}[\delta_k] = \mathbb{P}(\delta_k = 1) = p_k$ and variance $Var(\delta_k) = p_k(1-p_k)$, and the expected overperformance becomes:

$$\mathbb{E}[d_k] = \rho_{\delta,\xi}\sqrt{p_k(1-p_k)}\sigma_\xi + (p_k - 1)\mathbb{E}[\xi] - \varphi \tag{3}$$

Conversely, when short trades are allowed and functional, δ_k can take either 1 (with probability p_k) or -1 (with probability $1 - p_k$).[4] In this case, δ_k is a *Bernoulli*-like random variable ($(\delta_k + 1)/2$ is *Bernoulli*), now having mean $\mathbb{E}[\delta_k] = 2p_k - 1$ and variance $Var(\delta_k) = 4p_k(1 - p_k)$; the expected overperformance becomes:

$$\mathbb{E}[d_k] = 2\rho_{\delta,\xi}\sqrt{p_k(1-p_k)}\sigma_\xi + (2p_k - 1)\mathbb{E}[\xi] - \varphi \tag{4}$$

At this point, several notes are in order. First, p_k can be interpreted as the *bullish tendency* of trading rule k, i.e. its inclination to trade with (to follow) the market, while $(1-p_k)$ as the *bearish tendency*, i.e. its inclination to trade against (to stay out of) the market. The buy-and-hold rule has $p_0 = 1$ and $1 - p_0 = 0$, while all active trading rules have $p_k < 1$ and $1 - p_k > 0$.

Second, and very important for our discussion, benchmarking against the buy-and-hold strategy induces an asymmetry in the performance measure in asset return space: Positive excess returns can only be achieved when correctly predicting price declines (negative returns), while negative

[4]We prefer this specification, which implicitly assumes that trading rules are always active in the market, because it constitutes a limiting case that better highlights our point. Additionally considering zero as a signal would dampen the volatility of trading signals but would otherwise not change the conclusion of the analysis.

excess returns can only be achieved when incorrectly predicting price increases (positive returns). In other words, only the timing of predictions that go against the market can change TTR performance. Going into the details, from Eq. 3 we get (similar statements can be made based on Eq. 4) $\rho_{\delta,\xi} = 0 \implies \mathbb{E}[d_k] > 0 \iff \mathbb{E}[\xi] < -\frac{\varphi}{(1-p_k)} < 0$, which shows that even trading rules with zero predictive accuracy can benefit traders when used on financial assets or in periods with a sufficiently low, negative expected return, this being impossible when expected returns are positive. Also, $\frac{\partial \mathbb{E}[d_k]}{\partial \mathbb{E}[\xi]}\Big|_{\rho_{\delta,\xi}=0} = -(1-p_k) < 0$ and $\frac{\partial \mathbb{E}[d_k]}{\partial (1-p_k)}\Big|_{\rho_{\delta,\xi}=0} = -\mathbb{E}[\xi]$, which additionally shows that the benefits of bearish-inclined trading rules increase as their bearish tendency is stronger and/or as expected returns become more negative. This property of speculative trading rules should not be relevant for traders with long investment horizons interested in financial assets that have positive expected returns, such as stocks or bonds. However, (assuming time-varying expected returns) it should be important for traders with short investment horizons during periods of market crashes or for traders interested in financial assets with negative expected returns. Nevertheless, given limitations associated to applying MTPs in practice, it should be relevant for all empirical researchers, as we will next discuss.

Third, empirical tests evaluate the maximum expected excess performance obtained by any trading rule from a set of many candidates, but can only rely on finite-sample estimates of excess performance, i.e. $\hat{d}_k = \frac{1}{T}\sum_{t=1}^{T} \hat{d}_{k,t}$, where $T < \infty$ is the sample length. Thus, we can write the

following decomposition of the observed maximum average excess return:

$$\max_{k=1..K} \hat{d}_k = \max_{k=1..K} \underbrace{c\hat{\rho}_{\delta,\xi}\sqrt{\hat{p}_k(1-\hat{p}_k)}\hat{\sigma}_\xi}_{\text{predictive ability}} + \underbrace{(c\hat{p}_k - 1)\hat{\xi}}_{\text{directional tendency}} - \underbrace{[\hat{\theta}(\hat{\phi} + \hat{\lambda}_t) + \hat{\sigma}_{\delta,\lambda}]}_{\text{trading style/cost penalty}} \tag{5}$$

where $c = 2$ when short trades are allowed and $c = 1$ otherwise. Evidently, increasing the size of the sample can reduce estimation error. However, our analysis shows that other characteristics of the data sample may independently impact MTP results, opening up the possibility to multiple forms of data snooping bias. Eq. 5 shows how the estimated maximum excess performance depends on testing conditions and gives us an intuition about how data snooping bias can arise in empirical tests. In particular, Eq. 5 shows that there are three independent sources of trading rule overperformance: (1) (in-sample) *predictive ability*, (2) (in-sample) *directional tendency*, and (3) (in-sample) *trading style*. Thus, lucky trading rule characteristics or test misspecification could give rise to data snooping bias in three independent ways:

1. Via a component that depends on the predictive accuracy (ability) of the best trading rule. When the predictive accuracy is economically significant, i.e. $\rho_{\delta,\xi} > 0$, then the rule may help traders earn systematic excess returns. However, investment professionals and researchers routinely mine financial prices in search of better alternatives, this increasing the chances of finding and using rules that correlate with asset returns purely by chance and not due to real, economically significant ability. In the limit, $\max_{k=1..K} \hat{\rho}_{\delta,\xi} \gg \max_{k=1..K} \rho_{\delta,\xi}$ as $K \to \infty$. In turn, a lucky predictive accuracy falsely inflates the estimated excess perfor-

mance, thus creating a *spurious correlation effect* that can bias test results. This effect is the major source of concern for the discussion on data snooping in the literature, even if not always explicitly identified as such, e.g., the seminal work of Brock et al. (1992) or the discussion in Sullivan et al. (1999). Eq. 5 further shows that the spurious correlation effect can be exacerbated when: (i) asset volatility increases, (ii) trading signal volatility increases, or (iii) researchers allow short trades when performing tests on assets (in markets) where they are not functional.

2. Via a component that depends on the tendency of the best rule to trade with or against the market. As previously noted, active trading rules benefit investors when asset prices tend to decrease, and are disadvantageous otherwise. Moreover, the benefits associated with using active trading rules in downward trending markets increase with their bearish tendency and/or with the negative magnitude of returns. However, prices for many classes of assets do not typically trend downwards and could actually have a long-run positive price drift, making bearish-inclined trading rules suboptimal in many cases. More importantly, and similar to spurious correlation, the increased performance obtained in a sample exhibiting a negative average return could be the result of luck and not skill, as extensive data snooping can uncover trading rules with a spurious bearish tendency. In the limit, $\hat{p}_m \to \hat{p}_\xi$ as $K \to \infty$, where $m = \arg\max_{k=1..K} \hat{d}_k$ and $\hat{p}_\xi = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\{\xi_t>0\}}$. This, in turn, may contribute to the aggregate data snooping bias via a *spurious bearish tendency effect*. Eq. 5 further shows that this effect can be ex-

acerbated when researchers allow trading rules to take short positions in markets where they are not functional.

3. Via a component that depends on a rule's trading style, but also on how rigorous are researchers when accounting for trading costs. In a frictionless environment, aggressive trading rules (i.e. ones that can be characterized by a high trading frequency) have a natural tendency to overperform more passive alternatives because they can better adapt to short-term trends. However, trading costs are important in real markets, and ignoring or underestimating them would artificially inflate the performance of trading rules[5], especially for those that trade more often. This, in turn, would lead to a *trading style/costs effect* that can further bias results. Eq. 5 further shows that this effect can be exacerbated when lucky trading rules have a preference for illiquidity, i.e. trade more in less liquid, more volatile periods. Thus, misspecifying trading costs especially overestimates the performance of lucky rules in thinly traded markets.

In the end, MTPs that evaluate the performance of speculative trading rules are defined using asymptotic considerations, but extensive data snooping efforts, the limitations associated with applying them in practice, and researcher intentional or unintentional flawed implementation strategies can lead to biased results. Here, we identify the sources of luck and false discoveries that are important when many trading rules are tested on the same historical data samples. How much does each factor contribute to the aggre-

[5]This is a well-known source of bias. For example, Fama (1965) provides an early example in this direction when discussing the findings of Alexander (1961);

gate data snooping bias? We investigate this question next.

3. Drivers of data snooping bias

3.1. Setup

We set out to find the breaking point of modern MTPs under extreme data snooping pressure. For this, we design and perform a large scale stress test, which relies on multiple Monte Carlo simulations that evaluate the interaction between testing conditions, characteristics of lucky trading rules, and data snooping bias. On the one hand, we account for extensive data snooping efforts by selecting lucky rules from a very large set of 688,740 TTRs, which are largely defined following Anghel (2021a).[6] On the other hand, we take account of possible market conditions by constructing and using multiple data samples with varying properties. The baseline data consists of 6x1,000=6,000 samples of randomly generated trading prices and volumes spanning one month (approximately 22 observations each), which are obtained from a discretized no-drift geometric Brownian Motion with volatility parameters $\sigma \in \{0.15, 0.20, 0.25, 0.30, 0.35, 0.40\}$.[7]

Testing a very large trading rule universe on very short (one month) data samples can be seen as going against MTP theory. While this is correct, we emphasize that our focus here is not to evaluate trading rules with inappropriately applied tests, but rather to evaluate the resilience of MPTs

[6] We use the 686,304 trading rules defined by Anghel (2021a), to which we add 2,435 rules constructed from the Trading Range Breakouts (TRB) method, which has been previously considered by Sullivan et al. (1999), and a passive rule that always stays out of the market.

[7] The data construction also follows Anghel (2021a, Section 4.1).

given empirical realities, which are substantially different compared to the theoretical conditions. In particular, our approach is meant to capture the spirit of current data snooping trends, with traders and researchers evaluating millions or even billions of trading rules (admittedly using larger samples). Working with short samples also comes with several benefits for the analysis. First, it greatly eases computational demands, which in turn enables us to significantly grow the number of test runs and better estimate MTP error rates. Second, it exposes MTPs to additional sources of (estimation) error, which in turn enables us to truly meet the stress test goal of finding their limits in empirical applications. Third, it leads to a small sample evaluation of modern MPTs, which covers a gap in the literature.[8] Also, it is directly relevant for papers that do consider shorter samples when evaluating time-varying TTR performance, market efficiency. Note that, despite the many advantages, we do not ignore possible concerns regarding small sample bias and set out to validate all important findings in robustness tests that rely on significantly larger samples.

Moving on, the evaluation procedure goes as follows. On each sample, in *Step 1* we estimate the performance of all trading rules and isolate the one that earns the highest excess return, i.e. the luckiest rule. Then, in *Step 2* we evaluate the statistical significance of the results using popular MTPs that are designed to control for data snooping, namely the Reality Check test of White (2000) and the Superior Predictive Ability test of Hansen (2005).

[8]The analysis of the small sample properties of econometric models and statistical tests is a very important area of study in econometrics. However, there is close to no evidence on the behavior of MTPs in small samples.

Both tests control for the Family-wise Error Rate (FWER), making them less powerful and more conservative compared to alternatives that control for (variants of) the False Discovery Rate, FDR (see Barras, Scaillet, and Wermers, 2010; Efron, 2012). However, they are also less prone to making Type I errors, which is the metric that we are interested in. Thus, our analysis can be interpreted as gauging the utmost conditions under which MTPs are expected to fail, setting a minimum threshold for more liberal testing approaches.

There are several differences in our analysis compared to similar endeavors in the related asset pricing literature, such as the work of Chordia et al. (2020). First, we focus on analyzing the luck of speculative trading rules when applied on individual time-series (assets) and do not form or test portfolios of assets. This bypasses the need to assume a factor model for the data generating process and implies that our results can be generalized to a wide set of financial asset classes. Second, we assume no dependence structure of the time series that we simulate, thus making any test null rejection a false discovery and enforcing the null hypothesis of no trading rule overperformance directly from the construction of the dataset. Third, because we investigate trading rules based on technical analysis, we also simulate other relevant market information besides Closing prices (returns), such as Open, High, and Low prices, alongside trading volumes.

There are also some important differences compared to the work of Anghel (2021a), on which our analysis builds upon. On the one hand, we compute and use data sample statistics to analyze the interaction between false discoveries and sample characteristics, such as the average return, standard de-

viation of returns, skewness, and kurtosis. Among others, this enables us to isolate the contribution that spurious bearish tendency has on the aggregate data snooping bias. On the other hand, we do not fix the testing conditions, instead varying them based on methodological choices that applied researchers would need to make. First, we run the tests with and without a fixed broker fee.[9] Second, we run the tests with and without liquidity/price impact costs.[10] Third, we run the tests with and without allowing short trades. Fourth, we run the tests with and without standardizing the statistic. When the statistic is not standardized, the test follows the procedure described by White (2000). Conversely, when the statistic is standardized, the test follows the procedure described by Hansen (2005), which additionally makes an asymptotic adjustment for the impact of poor performing, irrelevant trading rules. The Bootstrap-Bonferroni adjustment of Romano and Wolf (2018) (which is based on finite-sample considerations) is also examined, but similar to Anghel (2021a) we find no material differences compared to the adjustment of Hansen (2005), thus opting not to discuss it further.

Overall, there are 16 different methodological combinations that we test on all 6,000 samples, amounting to a total of 96,000 test runs. Comparing the results enables us to estimate the impact of various testing conditions on the characteristics of lucky trading rules and on the data snooping bias that they would introduce in empirical tests.

[9] A fee of 1% per round trip is considered, which should be sufficient to account for trading fees in most financial markets.

[10] When considering liquidity/price impact costs, we simulate trading at the least favorable daily prices, the High price for buy trades and the Low price for sell trades (otherwise, trades are simulated at the average daily price).

3.2. Results: Testing conditions and trading rule characteristics

We first analyze how the characteristics of lucky trading rules change with testing conditions, i.e. the results following *Step 1* of the evaluation procedure. Figure 1 shows the conditional distribution of each characteristic given possible methodological choices. We offer additional details on this relationship in Appendix A, where Figures A.1 trough A.4 report scatter plots showing how each characteristic relates to the first four moments of the sample return distribution, given each methodological choice. Table 1 reports the results of four linear regression models that evaluate how testing conditions drive the characteristics. Several results are worth noting.

[Fig. 1 about here.]

[Table 1 about here.]

First, we find that an inflated predictive accuracy ($\hat{\rho}_{\delta,\xi}$) is the most stable characteristic of lucky trading rules, as shown by the highly significant intercept ($\alpha = 0.4485$, $t_\alpha = 142.19$) and low explanatory power of the regression model ($R^2 = 0.063$). This result implies that spurious correlation−even though it increases with the number of trading rules being considered−is largely independent from other testing conditions. Nevertheless, we do find that predictive accuracy increases with volatility, showing that spurious correlation rises in more volatile markets. Also, considering trading fees, restricting short trades, and standardizing the test statistic can help reduce spurious correlation, but only marginally. Interestingly, we further find a weak positive influence of market returns on predictive accuracy, hinting that spurious correlation slightly increases in upward trending markets.

Second, we find that trading frequency ($\hat{\theta}$) is slightly more dependent on testing conditions ($R^2 = 0.364$). In particular, adjusting for trading costs significantly decreases the trading frequency of lucky rules (as expected), while considering short trades and standardizing the test statistic have a mostly positive, although weaker influence. In the latter cases, the impacts depend on the properties of the data sample: For the most part, trading frequency tends to decrease with average returns and increase with volatility.

Third, we find that the directional tendency ($\hat{p}$) of lucky trading rules is highly dependent on testing conditions ($R^2 = 0.573$). Most notably, directional tendency is strongly determined by market returns: Lucky rules have a bearish tendency when returns decrease (are negative) and a bullish tendency when returns increase (are positive), this showing that spurious directional tendency is an conspicuous factor that could potentially influence test results. Adding trading fees and liquidity costs increase the impact of average returns on directional tendency by about 20% each[11], showing that trading costs can potentially have an adverse impact on test results trough this channel. We also find that restricting short trades and standardizing the test statistic decrease the impact of average returns on directional tendency by about -10% each, but that making these methodological choices generally favors rules that trade with the market more often.

Fourth, we find that the average (annualized) excess return ($\hat{d}$) of lucky trading rules is to a very large extent explained by testing conditions ($R^2 = 0.846$). On the one hand, sample volatility has a strong positive impact,

[11]This is estimated as the ratio between γ_{fee} or γ_{liq}, respectively, and γ (which is the coefficient showing the base effect of market returns on market tendency).

showing that the *spurious correlation effect* is exacerbated in more volatile markets, just as expected. Nevertheless, this effect can be decreased by restricting short trades (-31%), standardizing the test statistic (-14%), adjusting for liquidity costs (-9%), and adjusting for trading fees (-3%). On the other hand, and more importantly, we find a strong negative correlation with realized market returns, which shows that the *spurious bearish tendency effect* is the main factor that inflates the overperformance of speculative trading rules, especially when both long and short trades are considered (restricting short trades decreases the effect by -41%) and the test statistic is not standardized (standardizing the statistic decreases the effect by -12%). Interestingly, while trading fees have a direct negative influence on excess returns, they are not sufficient in eliminating the combined influence of spurious directional tendency and correlation. Also, liquidity costs do not have a direct impact on excess returns, instead only being useful in reducing the impact of volatility.

Overall, the results show that the characteristics of lucky trading rules and their estimated trading performance following extensive data snooping efforts are influenced by testing conditions in complex, non-trivial ways. In general, we find that spurious correlation is a pervasive characteristic that falsely increases trading rule performance regardless of testing conditions. Nevertheless, a bearish tendency in samples with low (negative) average returns seems to be the main factor that inflates average excess returns. Controlling for trading costs and restrictions on short selling positions is important for handling luck, but only to a limited extend. Interestingly, adding trading costs favors trading rules with a bearish directional tendency and could thus

have a negative influence on test results in some situations.

3.3. Results: Methodological choices and data snooping bias

We go on to *Step 2* and estimate Type I error frequencies for the RC (White, 2000) and SPA (Hansen, 2005) tests. These MTPs are designed to control for data snooping, but the asymptotically-valid theory may fail when faced with finite-sample limitations and/or unreasonable empirical implementation strategies.

We start by performing single-hypothesis tests, i.e. testing the luckiest trading rule in each sample, and then move on to multiple-hypothesis tests by gradually adding the other rules. For each of the 96,000 test runs, we analyze 21 increasingly larger sets of trading rules, resulting in a total of 2,016,000 individual tests. For each test, we estimate the distribution of the statistic as suggested by White (2000) and Hansen (2005), i.e. by bootstrapping from trading rule excess returns using the Stationary Bootstrap of Politis and Romano (1994) with 1,000 iterations. From this, we estimate critical values and test p-values. Finally, we estimate the empirical FWER (i.e., the aggregate proportion of false discoveries) at the standard 5% significance level, aggregating by the various testing conditions. A summary of the results is reported in Table 2.

[Table 2 about here.]

Overall, we find that between 21% and 78% (48% on average) of single hypothesis tests (i.e., tests that only evaluate the luckiest trading rule and do not account for other rules) and between 1% and 12% (2.4% on average) of full-scale multiple hypothesis tests (i.e., tests that account for all

trading rules) can falsely reject a true null hypothesis, depending on testing conditions. The differences in error rates between the single and multiple hypothesis tests, which range between 20.1 and 66.5 percentage points (43.1 on average), provide estimates for the amount of false discoveries due to failing to account for combined data snooping efforts. These results support the findings of Anghel (2021a), but extends them by additionally showing that data snooping bias from misspecifying trading rule universes significantly varies with testing conditions. Table 3 shows exactly how by reporting the average contribution of each methodological choice to the aggregate bias.

[Table 3 about here.]

First, ignoring trading fees and liquidity costs generates false discoveries in 3.8%-18.8% and 3.2%-14.4% of tests, respectively. On average, this accounts for 44.9% and 36.8% of the total bias. Also, ignoring trading costs leads to significant size distortions, as error rates surpass the 5% threshold in all associated tests, including after all 688,740 trading rules have been accounted for. This shows that incorporating trading costs is a minimum requirement for adequately controlling data snooping and assuring the consistency of tests in finite data samples.

Second, the contribution of restricting short trades on the aggregate bias is low (2.6% on average), and also depends on other testing conditions. In particular, restricting short trades increases false discoveries when few rules are tested together, but decreases them after accounting for all trading rules. This implies that full-scale multiple-hypothesis tests are better at controlling data snooping in markets where short trades are functional.

Third, standardizing the test statistic is generally detrimental to the analysis because it increases false discoveries by between 1.2 and 25.1 percentage points, accounting for 4.3% to 60.3% of the total bias. Nevertheless, upon closer inspection, we find that standardizing the test statistic is detrimental only when short trades are restricted, but is beneficial (reduces false discoveries) otherwise. This implies that standardizing the test statistic is a better testing strategy when the volatility of trading signals is expected to be high, such as in markets where short trades are functional, but could otherwise escalate the data snooping bias. This finding explains the results in Anghel (2021a), which only considered long trades and found RC tests to be better at controlling data snooping compared with SPA tests. Here, we show that this is not necessarily the case and, instead, researchers should choose between the two testing strategies depending on market conditions.

Fourth, we estimate that a t-ratio of 3.24 is required on average to adequately control for data snooping in tests performed at a 5% significance level (2.63 and 4.77 for the 10% and 1% levels, respectively). This may rise to 4.17 with some inappropriate methodological choices, and even to 4.37 when trading costs are ignored.

3.4. Results: Data sample properties and data snooping bias

We continue the investigation by analyzing if data sample properties also play a role in shaping data snooping bias. Indeed, our main novel finding is that false discoveries are overwhelmingly concentrated on samples with negative average returns, irrespective of methodological choices. The results are reported in Table 4: Panel A shows aggregate error rates in tests performed on samples with a positive average return; Panels B shows aggregate

error rates in tests performed on samples with a negative average return; and Panel C shows the proportion of false discoveries that arise in samples where average returns are negative relative to the total number of false discoveries.[12]

[Table 4 about here.]

We find that between 41% and 97% (71% on average) of single hypothesis tests performed on downward-trending samples can falsely reject a true null hypothesis, representing 71% of the total. This result implies that the *spurious bearish tendency effect* dominates the *spurious correlation effect* right from the onset in terms of its potential to generate data snooping bias. More importantly, as the number of trading rules that are considered increases (to control for data snooping), so does the relative contribution of the former to the aggregate bias. After controlling for all data snooping efforts in full-scale multiple hypothesis tests, between 2% and 24% (4.7% on average) of tests performed on downward-trending samples lead to false discoveries, representing almost 100% of the total. This holds for all methodological combinations and implies that accounting for data snooping efforts is effective at controlling spurious correlation, but not spurious bearish tendency. In the end, this finding shows that modern MTPs can be severely limited in specific empirical applications.

The results also show a potential remedy for this limitation. Specifically,

[12]We also analyze if test results significantly differ in samples with: (1) high vs. low Standard Deviation of returns; (2) positive vs. negative Skewness; and (3) positive vs. negative Excess Kurtosis. We find no significant differences in these cases, except a slight asymmetry by skewness (false discoveries occur at higher rates when sample Skewness is positive) when a standardized test statistic is used. For the sake of brevity, we defer a detailed analysis of this result to future work.

a t-ratio of 3.85 can be used to control for data snooping bias at a 5% significance level in bearish-leaning samples (3.21 and 5.33 for the 10% and 1% levels, respectively), this rising to 4.91 with some poor methodological choices, and even to 5.10 when trading costs are ignored. Note that the higher significance thresholds constitute a two-fold increase compared to the ones that are required in bullish-leaning samples, a results that can be observed both on average, as well as for each methodological combination. Recent evidence from the asset pricing literature (e.g. Harvey et al., 2016) show that considering the standard 1.96 threshold (1.64 and 2.57 for the 10% and 1% levels, respectively) makes it highly likely that positive discoveries with respect to the overperformance of trading strategies are false, and that raising the statistical threshold to well above 3 is required to account for luck. Here, we provide evidence that this is also true when examining TTR overperformance. However, our results go one step further and provide novel evidence showing that size distortions of MTPs are heterogeneous, as they also depend on the empirical distribution of asset (market) returns. This implies that sticking to the same (although higher) significance threshold and ignoring market conditions is not the optimal testing approach in samples where asset prices trend downward. Instead, the latter conditions demand an asymmetric treatment of test significance thresholds, or improvements in test design as to control for this asymmetry.

[Table 5 about here.]

We further check for possible interactions between different sample properties by analyzing error rates on bivariate sample sorts. Table 5 reports error

rates when the results are grouped by sample average return and standard deviation of returns deciles, which turn out to present the most significant variations. We find that a low volatility increases size distortions due to the spurious bearish tendency effect. In particular, the proportion of false discoveries even for full-scale multiple hypothesis tests always exceeds the 5% level when the average return is in the first decile ($\hat{\xi} \leq -123.48\%$ per year), and increases as volatility decreases. However, positive distortions are also significant for the second average return decile ($\hat{\xi} \leq -76.26\%$ per year) when the standard deviation of returns is lower than 26.38%, and even for the third return decile ($\hat{\xi} \leq -47.58\%$ per year) when the standard deviation of returns is lower than 15.05%.

3.5. Results: Testing conditions and data snooping bias

We complement the analysis by estimating several regression models that show how methodological choices and sample properties interact to generate false discoveries. The results are reported in Table 6. We note that the combination of four methodological choices and two data sample properties explains false discoveries to a statistically and economically significant extent, in both single and multiple hypothesis tests.[13] This is evident when observing that: (i) almost all estimated coefficients are statistically significant, most of the time at the 1% confidence level; and (ii) the R^2 in all regressions ranges between 0.512 and 0.702. We focus the analysis on the logit models, which show how the probability of making false discoveries varies with testing conditions. Several interesting findings are worth noting.

[13] Adding sample skewness and kurtosis to the regressions does not qualitatively change, improve the results.

[Table 6 about here.]

First, methodological choices have a direct and significant impact on test outcomes: Deducting trading costs always brings net benefits to the analysis, as it has a consistent negative impact on false discoveries, while restricting short trades and standardizing the test statistic are beneficial only in full-scale multiple hypothesis tests, after accounting for all data snooping efforts.

Second, data sample properties also have a direct and significant impact on test outcomes. On the one hand, average returns always have a negative influence on the probability of falsely rejecting the null, supporting previous observations on the role of the *spurious bearish tendency effect* in shaping data snooping bias. Moreover, the impact of average returns is stronger in full-scale multiple hypothesis tests, complementing previous evidence showing that the *spurious bearish tendency effect* is solely responsible for any residual errors after accounting for data snooping efforts. On the other hand, the sample standard deviation also plays a role in shaping test outcomes, having a positive (negative) impact on the probability of rejecting the null in single (multiple) hypothesis tests.[14] The sign difference between single and multiple hypothesis tests implies that the magnitude of the *spurious correlation effect* increases with volatility, but that accounting for data

[14]Our results seemingly contradict Anghel (2021a), which finds a positive relationship between market volatility and false discoveries in multiple hypothesis tests. However, the difference can be easily explained by how our results generalize. Indeed, we also find false discoveries to increase with volatility when performing tests with trading costs enabled and short trades disabled, such as in Anghel (2021a). However, we find that the overall sign of the relationship is negative after considering the effects of other methodological choices and sample properties. Overall, our results are broader and highlight that methodological choices should be tailored to the expected market conditions.

snooping efforts in multiple-hypothesis tests is efficient at eliminating its influence. Taken together, the results show that the risk of falsely rejecting the null is highly persistent in downward trending, less volatile samples, even after seemingly accounting for data snooping efforts.

Third, methodological choices also have an indirect impact on test outcomes by shaping the way in which test results interact with sample characteristics. In general, restricting short trades and standardizing the test statistic weaken the effects of sample properties on test outcomes, while, interestingly, subtracting trading costs strengthens them. This latter result shows the complex inner working of data snooping. Although subtracting trading costs has a net negative impact on false discoveries, it also favors trading rule characteristics that increase data snooping bias, e.g. it has a negative influence on the bullish tendency of lucky trading rules, which in turn exacerbates the *spurious bearish tendency effect*.

3.6. Robustness check: Longer data samples

To alleviate potential concerns that our results are driven by large estimation errors due to using small data samples, we repeat the analysis on 6,000 longer samples, which span for 1 year (approximately 261 observations) or 4 years (approximately 1044 observations). For brevity, we fix the methodological choices and only perform tests in which both trading fees and liquidity costs are considered, short trades are disabled, and a simple performance measure (the RC test) is used. The results are reported in Tables 7 (1 year samples) and 8 (4 year samples), showing the previous conclusions hold when longer samples are used: Size distortions remain significantly positive for the first average return decile, i.e. when $\hat{\xi} \leq -35.9\%$ per year for 1 year samples

and $\hat{\xi} \leq -17.8\%$ per year for 4 year samples; while negatively correlating with volatility.

[Table 7 about here.]

[Table 8 about here.]

We do notice a slight decrease in overall error rates compared to the baseline results, potentially showing that using larger samples better controls for spurious bearish tendency and improves the size of multiple testing procedures. However, a closer look shows that the improvement is not due to lower estimation errors, but rather to a beneficial but unintended convergence in sample average returns to their population value (which is $\mu = 0$), thus decreasing the potential benefit of and indirectly hindering the spurious bearish tendency effect. We specifically check this hypothesis by running additional tests on 6x100 even longer, 10 year samples (containing approximately 2,610 observations each) that are drawn from a Geometric Brownian Motion (GBM) with the same volatility parameters as before, but now having a drift of $\mu = -0.2$.

[Table 9 about here.]

Table 9 reports the results, confirming that false discoveries abound even in very long data samples when average (expected) returns are negative, and that increasing the sample length does not directly benefit the analysis. Overall, the results in this Section show that data snooping bias from spurious bearish tendency is very persistent when sample average returns are negative, irrespective of sample length. We do note, however, that the interaction

between sample average return and sample length does have an influence on test results: For the same average return, false discoveries increase on longer samples. This implies that spurious bearish tendency has a higher impact when the sample length increases, and that data snooping bias can persist in tests conducted on financial assets that have experienced minor price declines over long periods of time.

3.7. Robustness check 2: Other testing procedures

Up to now, our analysis has centered on tests that control for the Family-wise Error Rate, which can be traced back to the seminal contribution of White (2000). Nevertheless, FDR tests can also be subject to limitations as, e.g., Andrikogiannopoulou and Papakonstantinou (2019) and Barras, Scaillet, and Wermers (2021) recently discuss. To verify if this is indeed true in our case, we further analyze the robustness of results to changes in the testing procedure, i.e. to using the more powerful but less stringent FDR control, which has become fairly popular in the financial economics literature (see Harvey et al., 2020, and references therein). Here, we consider the pFDR test of Storey (2002). The testing conditions follow the ones defined in Section 3.6, except that we focus on 1 year samples.

[Table 10 about here.]

The results, which are reported in Table 10, show that a similar negative relationship between data snooping bias and sample average return exists when controlling for pFDR, with size distortions being especially significant for the first two return deciles. The difference in this case is that the distortions are generally larger, and do not monotonically decrease with volatility,

instead having an U shape in volatility space for low return deciles and an *inverted-U* shape for middle and higher return deciles. Also, size distortions in pFDR tests remain significant for all return deciles when volatility falls in the middle deciles. Overall, the results imply that data snooping bias from the spurious bearish tendency of lucky trading rules is even more significant when more liberal MTPs are used, in particular (p)FDR-based ones. This is expected, as the analysis centered on tests that control for FWER is the most conservative and basically provides minimum conditions for when MTPs fail under extreme data snooping pressure.

4. An empirical analysis of trading rule overperformance

4.1. The cryptocurrency market

We build on the Monte Carlo simulation and perform an empirical investigation of possible spurious effects and the luck-adjusted overperformance of TTRs in real financial markets. First, we focus on the cryptocurrency market. Besides the recent surge in academic interest towards it (Jiang, Li, and Wang, 2021), the study of cryptocurrencies is relevant in our context because they are seemingly "less efficient" compared to more traditional financial assets. In particular, existing evidence shows that cryptocurrency returns are more predictable compared to returns in other markets (Zhang, Wang, Li, and Shen, 2018; Al-Yahyaee, Mensi, and Yoon, 2018; Sensoy, 2019), while TTRs seem to help investors earn statistically and economically significant "abnormal" returns (Corbet et al., 2019; Grobys et al., 2020; Fischer, Krauss, and Deinert, 2019; Hudson and Urquhart, 2021).

However, the recent surge in interest toward this market and the general

lack of control for luck may lead to results that are plagued by data snooping bias. Of particular interest to our discussion are papers such as Gerritsen, Bouri, Ramezanifar, and Roubaud (2020), who observe that the added value of trading rules depends on market conditions (periods of overperformance are related to strongly trending markets), but implicitly attribute this finding to a superior ability of trading rules without considering possible spurious effects. Are trading rules really profitable in the cryptocurrency market after adjusting for luck?

Recently, Anghel (2021b) used 67,480 trading rules based on technical analysis and 5 rules based on machine learning algorithms to show that statistically and economically significant excess returns can be hardly achieved after controlling for market frictions and data snooping. Here, we greatly expand the analysis by testing our set of 688,740 trading rules, while also decomposing null rejections by data sample properties to evaluate the role played by spurious bearish tendency in shaping the results. More importantly, we not only test for statistical significance, but also evaluate economic significance by comparing test results with the results obtained in the Monte Carlo simulations (reported in Section 3), which we consider as a benchmark. If the rates of positive discoveries (test null rejections) would exceed what we expect at the bounds of randomness, than trading rules can be deemed as having a superior forecasting ability.

In the empirical exercise we fix the methodological choices, tailoring them to the specific conditions encountered by traders. First, trading fees for cryptocurrencies vary by market/broker but can reach as low as 0%, so we decide not to consider them. Instead, liquidity and price impact costs should

be important in a relatively young and thinly traded market, so we decide to incorporate them. Second, short trades are possible, but only on some exchanges and mostly for top-tier pairs. In general, we assess that short trades cannot currently be considered as fully operational in the cryptocurrency market, so we decide to ignore them.[15] Finally, we choose to employ both a simple and a standardized performance measure (test statistic). The cryptocurrency market is on average more volatile compared to its competitors, which suggests that a standardized statistic should be used in principle. However, because we only consider long positions, we expect that a simple statistic would better control for false discoveries.

The data sample is obtained from Anghel (2021b) and consists of all available trading histories for 861 cryptocurrencies collected on February 10, 2020 from www.coinmarketcap.com. For each cryptoasset we measure the excess performance obtained by all trading rules in the extended set. Then, we split the data by calendar month and perform RC and SPA tests on monthly subsamples of at least 22 observations. As before, 21 different statistical tests are performed on each subsample by estimating the distribution of excess returns using 21 increasingly larger sets of trading rules. In total, we base our analysis on 1,344,588 individual test results obtained on 32,014 subsamples.

[Table 11 about here.]

[Table 12 about here.]

[15]Some papers go in the opposite direction. For example, Fischer et al. (2019) consider short trades but explicitly acknowledged that short-selling constraints may constitute a limit of their analysis.

Table 11 reports a decomposition of average RC and SPA null rejections by average sample return and standard deviation of returns deciles[16], while a summary of the results obtained in this empirical exercise and a comparison with the benchmark results are reported in Table 12. One the one hand, we find that null rejections are concentrated in the first return decile and also increase with lower volatility, including for multiple hypothesis tests that control for all trading rules from which best-performing ones are extracted (see Panel B in Table 11). This constitutes evidence that the spurious bearish tendency effect is important for shaping the results and implies that we should be more conservative when inferring economic significance.

On the other hand, we find that the tests performed in the cryptocurrency market reject their null hypotheses 50%-90% less often compared to the benchmark. A similar result is obtained on the subset of samples exhibiting negative average returns, implying that TTRs have no economic relevance in periods of declining cryptocurrency prices. Interestingly, testing few rules–including using single hypothesis tests–results in more null rejections on bearish samples compared with bullish samples, which is consistent with our simulation results. However, increasing the number of tested rules inverses the asymmetry and results in more null rejections on bullish samples. This implies that false discoveries arising from the spurious bearish reference effect are eliminated after adjusting for data snooping in the cryptocurrency market, while remaining null rejections could plausibly be linked with a statistically significant superior predictive accuracy of some trading rules when the market is trending upward.

[16] Analyzing the RC and SPA test separately points toward exactly the same conclusions.

A detailed analysis reveals 97 instances when the RC null hypothesis is rejected on samples with positive average returns, which constitutes 0.7% of all bullish samples (0.3% of all samples). This is higher than the proportion estimated in the simulation exercise using randomly generated data, which was 0%. Similarly, there are 146 instances when the SPA null is rejected on samples with positive average returns, which constitutes 1.1% of all bullish samples (0.4% of all samples), again being higher than the 0.2% baseline result. Null rejections decline when testing more rules, implying that a substantial amount of false discoveries due to the spurious correlation effect can be eliminated when controlling for data snooping efforts. However, the differences of 0.3% for the RC test and 0.9% for the SPA test constitute evidence in favor of TTRs being capable of successfully timing cryptocurrency prices and earning statistically significant excess returns in some periods of increasing prices. Nevertheless, we consider a 0.3%-0.9% success rate in bullish periods (representing less than 0.37% of all analyzed periods) not to be economically significant, as it would not encourage investors to use such trading rules to make real investment decisions. As a result, and taking into account the complete set of results, we conclude that TTRs cannot be considered as having an economically significant superior predictive ability, and that the the cryptocurrency market is very close to informationally efficient in the sense proposed by Timmermann and Granger (2004). This conclusion contradicts many of the previous findings reported in the literature, implying that data snooping is the main factor driving earlier result.

4.2. The stock market

We perform a complementary empirical exercise using data collected from the stock market, which enables us to apply tests on longer data samples with completely different characteristics. The literature on the profitability of trading rules in this market goes back for a longer period, but is generally subject to the same limitations (see Anghel, 2021a, and references therein). Again, of particular interest for our discussion are papers such as Fernandez-Rodrıguez, Gonzalez-Martel, and Sosvilla-Rivero (2000), who conclude that TTRs are superior to the buy-and-hold strategy in bear markets (being inferior in bull markets), but do not investigate further.

For this analysis, the data is obtained from Anghel (2021a) and consists of trading histories for 2,426 stocks listed in 77 markets (representing 84 countries) around the world. The testing procedure is similar to what was previously used for the cryptocurrency market, except that: (i) we perform statistical tests on 1 year subsamples averaging 231 observations (we exclude subsamples with less than 65 data points), or on full samples averaging 3,310 observations; (ii) we adjust excess returns with a fixed broker fee of 0.5% per transaction; (iii) we only use a simple test statistic (the RC test); and (iv) we only test the complete set of trading rules, thus always handling for combined data snooping efforts.

[Table 13 about here.]

Table 13 reports the decomposition of RC null rejections by average sample return and standard deviation of returns deciles. When testing 1 year subsamples, we find that 0.37% (96/34,887) of tests reject their null hypothesis, but that null rejections are concentrated in the first return decile and

largely increase with lower volatility. As before, this points to data snooping bias from spurious bearish tendency as being the main factor that drives the results, and that TTRs have no economically significant predictive ability when used on the stock market over short prediction horizons. However, the results obtained on the extended samples are qualitatively different: Now, 1.73% (42/2,426) of tests reject their null hypothesis, and there is no clear connection between sample characteristics and null rejection rates. This points towards the idea that data snooping is not an issue and that TTRs may some merit in timing stock returns when used over very long investment horizons.

On the one hand, the two results validate our previously stated hypothesis, i.e. spurious bearish tendency is not a problem for traders with long investment horizons interested in financial assets that have positive expected returns but should be important for traders with short investment horizons during periods of downward trending prices, or for researchers that study market efficiency using samples that have a bearish tendency. On the other hand, regarding the ability of TTRs for timing the stock market, the results points that they are still not economically superior to the buy and hold strategy. The first problem is that traders and investment professionals only use TTRs over short investment horizons, of up to 6 months (see Menkhoff, 2010, and related evidence). The second problem, similar to the analysis on cryptocurrencies, is that the proportion of null rejection can hardly be considered as economically significant, especially since (when looking in more detail) they are overwhelmingly concentrated on stocks listed in very small markets, which can be associated with greater market frictions compared

to what we have considered in this analysis. Overall, we can conclude that speculative trading rules have next to no economic relevance for traders in the stock market after adjusting for asymmetric luck, thus adding to the growing body of evidence in this direction (e.g. Sullivan et al., 1999; Taylor, 2014; Anghel, 2021a).

4.3. The FX market

Finally, we take a quick look at the foreign exchange market, in which assets (exchange rates) do not have a clear positive price drift. Instead, as predicted by the Uncovered Interest Rate Parity (UIP) condition, they may have a positive or a negative drift, depending on the interest rate differential between the two quoted currencies. For this analysis, we collect from Bloomberg all available trading histories for 9 major, minor and exotic currency pairs, i.e. AUDCAD, CADSEK, DKKZAR, EURAUD, GBPCHF, JPYSGD, JPYZAR, NZDZAR, and SGDHKD. The data series range from as early as January 4, 1971 to October 28, 2021. The testing procedure is identical to what was previously used for the stock market, except that (i) we adjust excess returns with a lower broker fee of 0.2% per transaction; (ii) we consider subsamples that span 4 years; (iii) we enable short trades, as they are functional in this market; and (iv) we also consider a standardized test statistic (the SPA test) to evaluate results, as it should be more reliable given the previous adjustment.

[Table 14 about here.]

Table 14 reports the results, clearly showing that null rejections are concentrated on bearish-leaning data samples. This result offers additional val-

idation to our earlier conclusions. First, a lucky bearish tendency of TTRs causes data snooping bias when tests are performed on bearish-leaning samples, irrespective of sample length. Third, TTRs do not have an economically significant superior forecasting ability when applied in the FX market. Because this conclusion disagrees with many of the previous findings reported in the literature, it implies that data snooping is a very important factor that drives earlier result. Third, when evaluating the superior predictive ability of forecasting models and associated trading rules using MTPs, researchers should account for the characteristics of the data sample that they use and, more generally, for the characteristics of the financial assets that they investigate.

5. Conclusions

While some guidelines exist on how to design and conduct a relevant empirical exercise when examining the profitability of speculative trading strategies, especially in the era of machine learning (see Arnott, Harvey, and Markowitz, 2019), applied researchers often go their own subjective way. The lack of evidence regarding how/when each testing condition can lead to false discoveries does not encourage a homogeneous approach. This increases the risk that positive findings reported in the literature may be biased due to data snooping. In this paper, we provide novel evidence that can help alleviate this problem.

First, we show that incorporating trading costs into the analysis is a must in order to assure finite sample control of MTP aggregate error rates. Second, short trading restrictions should be adequately enforced, otherwise

they might falsely inflate trading rule overperformance and might lead to false discoveries. When short trades are used, or when the signal volatility of trading rules is expected to be high, standardizing the performance measure (test statistic) would result in fewer Type I errors. Third, researchers must control as best as possible for all data snooping efforts, regardless of other testing conditions, in order to eliminate/reduce the bias induced by spurious correlation (which is a function of data snooping efforts but otherwise arises independently from other testing conditions) and bearish tendency (which arises in downward trending markets).

Nevertheless, our main novel finding is that spurious correlation has a small contribution to the aggregate data snooping bias, while the spurious bearish tendency of lucky trading rules is the main driver of false discoveries by a significant margin, irrespective of other testing conditions. Moreover, the spurious bearish tendency effect cannot be eliminated in multiple hypothesis tests and is solely responsible for the bias remaining after all of the necessary, previously stated methodological precautions have been taken. Even increasing the sample length does not make a difference when the average sample return remains negative. In the end, the probability of making false discoveries is non-negligible and MTPs fail under extreme data snooping efforts in downward trending, less volatile markets. In such cases, applied researchers must increase the statistical significance threshold and should be more prudent when interpreting results. Our findings thus support and extend recent evidence in the asset pricing literature (e.g. Harvey et al., 2020; Chordia et al., 2020) showing that false discoveries may abound, and add to the conclusion that accounting for lucky trading strategies requires raising

the statistical significance threshold to well above $t = 3$. Overall, our results show that a two-fold increase in statistical significance is required for tests conducted on bearish samples, compared to bullish samples. In particular, we estimate that a $t = 3.85$ is required (on average) to assure finite-sample control of MTP error rates at the 5% level in bearish-leaning samples, compared to $t = 3.24$ that is needed in all samples and $t = 1.96$ that is needed in bullish-leaning samples. Note that these baseline thresholds should increase with poor methodological choices and unfavorable sample conditions.

We also perform an empirical exercise to demonstrate how our results can be applied in tests examining the profitability of speculative trading rules in real financial markets. We focus on the emerging cryptocurrency market and the more mature stock and foreign exchange markets, and tune our methodological choices to adequately account for data snooping efforts. The results show that the spurious bearish preference effect is a factor that can influence test results in a wide array of practical circumstances, such as choice of statistical test, financial asset, and sample length. Taking asymmetric data snooping bias into account, we find almost no evidence in favor of TTR economically significant overperformance: "Abnormal" profit opportunities are insignificant and likely unattainable to investors that use TTRs to time these markets, which can be regarded as being informationally efficient with respect to a very large class of speculative trading strategies.

Finally, we note that our analysis shows that the asymmetry in MTP error rates is induced by the way we typically measure the performance of trading rules, i.e. as the excess return over a benchmark strategy that is always engaged in the market (the buy-and-hold rule). As a result, we would expect

our analysis to be relevant not only to evaluating speculative trading rules in financial markets, but to all forecasting exercises where and equivalent relative performance metric is employed. Currently, no MTP is designed to handle for asymmetrically lucky forecasting performance and data snooping bias, which implies the need to develop such procedures in the future.

References

K. H. Al-Yahyaee, W. Mensi, and S.-M. Yoon. Efficiency, multifractality, and the long-memory property of the bitcoin market: A comparative analysis with stock, currency, and gold markets. *Finance Research Letters*, 27: 228–234, 2018.

S. S. Alexander. Price movements in speculative markets: Trends or random walks. *Industrial Management Review (pre-1986)*, 2(2):25–46, 1961.

A. Andrikogiannopoulou and F. Papakonstantinou. Reassessing false discoveries in mutual fund performance: Skill, luck, or lack of power? *The Journal of Finance*, 74(5):2667–2688, 2019.

D. G. Anghel. Data snooping bias in tests of the relative performance of multiple forecasting models. *Journal of Banking & Finance*, page 106113, 2021a.

D. G. Anghel. A reality check on trading rule performance in the cryptocurrency market: Machine learning vs. technical analysis. *Finance Research Letters*, 39:101655, 2021b.

R. Arnott, C. R. Harvey, and H. Markowitz. A backtesting protocol in the era of machine learning. *The Journal of Financial Data Science*, 1(1): 64–74, 2019.

P. Bajgrowicz and O. Scaillet. Technical trading revisited: False discoveries, persistence tests, and transaction costs. *Journal of Financial Economics*, 106(3):473–491, 2012.

L. Barras, O. Scaillet, and R. Wermers. False discoveries in mutual fund performance: Measuring luck in estimated alphas. *The Journal of Finance*, 65(1):179–216, 2010.

L. Barras, O. Scaillet, and R. Wermers. Reassessing false discoveries in mutual fund performance: Skill, luck, or lack of power? a reply. *The Journal of Finance, Forthcoming*, 2021.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

W. Brock, J. Lakonishok, and B. LeBaron. Simple technical trading rules and the stochastic properties of stock returns. *The Journal of Finance*, 47 (5):1731–1764, 1992.

T. Chordia, A. Goyal, and A. Saretto. Anomalies and false rejections. *The Review of Financial Studies*, 33(5):2134–2179, 2020.

S. Corbet, V. Eraslan, B. Lucey, and A. Sensoy. The effectiveness of technical trading rules in cryptocurrency markets. *Finance Research Letters*, 31:32–37, 2019.

A. Cowles 3rd. Can stock market forecasters forecast? *Econometrica: Journal of the Econometric Society*, pages 309–324, 1933.

H. Dichtl, W. Drobetz, A. Neuhierl, and V.-S. Wendt. Data snooping in equity premium prediction. *International Journal of Forecasting*, 37(1): 72–94, 2021.

B. Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

E. F. Fama. The behavior of stock-market prices. *The Journal of Business*, 38(1):34–105, 1965.

F. Fernandez-Rodrıguez, C. Gonzalez-Martel, and S. Sosvilla-Rivero. On the profitability of technical trading rules based on artificial neural networks:: Evidence from the madrid stock market. *Economics Letters*, 69(1):89–94, 2000.

T. G. Fischer, C. Krauss, and A. Deinert. Statistical arbitrage in cryptocurrency markets. *Journal of Risk and Financial Management*, 12(1): 31, 2019.

D. F. Gerritsen, E. Bouri, E. Ramezanifar, and D. Roubaud. The profitability of technical trading rules in the bitcoin market. *Finance Research Letters*, 34:101263, 2020.

K. Grobys, S. Ahmed, and N. Sapkota. Technical trading rules in the cryptocurrency market. *Finance Research Letters*, 32:101396, 2020.

P. R. Hansen. A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4):365–380, 2005.

P. R. Hansen, A. Lunde, and J. M. Nason. The model confidence set. *Econometrica*, 79(2):453–497, 2011.

C. R. Harvey. Presidential address: The scientific outlook in financial economics. *The Journal of Finance*, 72(4):1399–1440, 2017.

C. R. Harvey, Y. Liu, and H. Zhu. ... and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68, 2016.

C. R. Harvey, Y. Liu, and A. Saretto. An evaluation of alternative multiple testing methods for finance applications. *The Review of Asset Pricing Studies*, 10(2):199–248, 2020.

P.-H. Hsu, Y.-C. Hsu, and C.-M. Kuan. Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias. *Journal of Empirical Finance*, 17(3):471–484, 2010.

R. Hudson and A. Urquhart. Technical trading and cryptocurrencies. *Annals of Operations Research*, 297(1):191–220, 2021.

S. Jiang, X. Li, and S. Wang. Exploring evolution trends in cryptocurrency study: From underlying technology to economic applications. *Finance Research Letters*, 38:101532, 2021.

J. T. Linnainmaa and M. R. Roberts. The history of the cross-section of stock returns. *The Review of Financial Studies*, 31(7):2606–2649, 2018.

R. D. McLean and J. Pontiff. Does academic research destroy stock return predictability? *The Journal of Finance*, 71(1):5–32, 2016.

L. Menkhoff. The use of technical analysis by fund managers: International evidence. *Journal of Banking & Finance*, 34(11):2573–2586, 2010.

J. P. Romano and M. Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282, 2005.

J. P. Romano and M. Wolf. Multiple testing of one-sided hypotheses: combining bonferroni and the bootstrap. In *International Conference of the Thailand Econometrics Society*, pages 78–94. Springer, 2018.

A. Sensoy. The inefficiency of bitcoin revisited: A high-frequency analysis with alternative currencies. *Finance Research Letters*, 28:68–73, 2019.

J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

R. Sullivan, A. Timmermann, and H. White. Data-snooping, technical trading rule performance, and the bootstrap. *The Journal of Finance*, 54(5): 1647–1691, 1999.

N. Taylor. The rise and fall of technical trading rule success. *Journal of Banking & Finance*, 40:286–302, 2014.

A. Timmermann and C. W. Granger. Efficient market hypothesis and forecasting. *International Journal of Forecasting*, 20(1):15–27, 2004.

H. White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000.

N. Zarrabi, S. Snaith, and J. Coakley. Fx technical trading rules can be profitable sometimes! *International Review of Financial Analysis*, 49:113–127, 2017.

W. Zhang, P. Wang, X. Li, and D. Shen. The inefficiency of cryptocurrency and its cross-correlation with dow jones industrial average. *Physica A: Statistical Mechanics and its Applications*, 510:658–670, 2018.

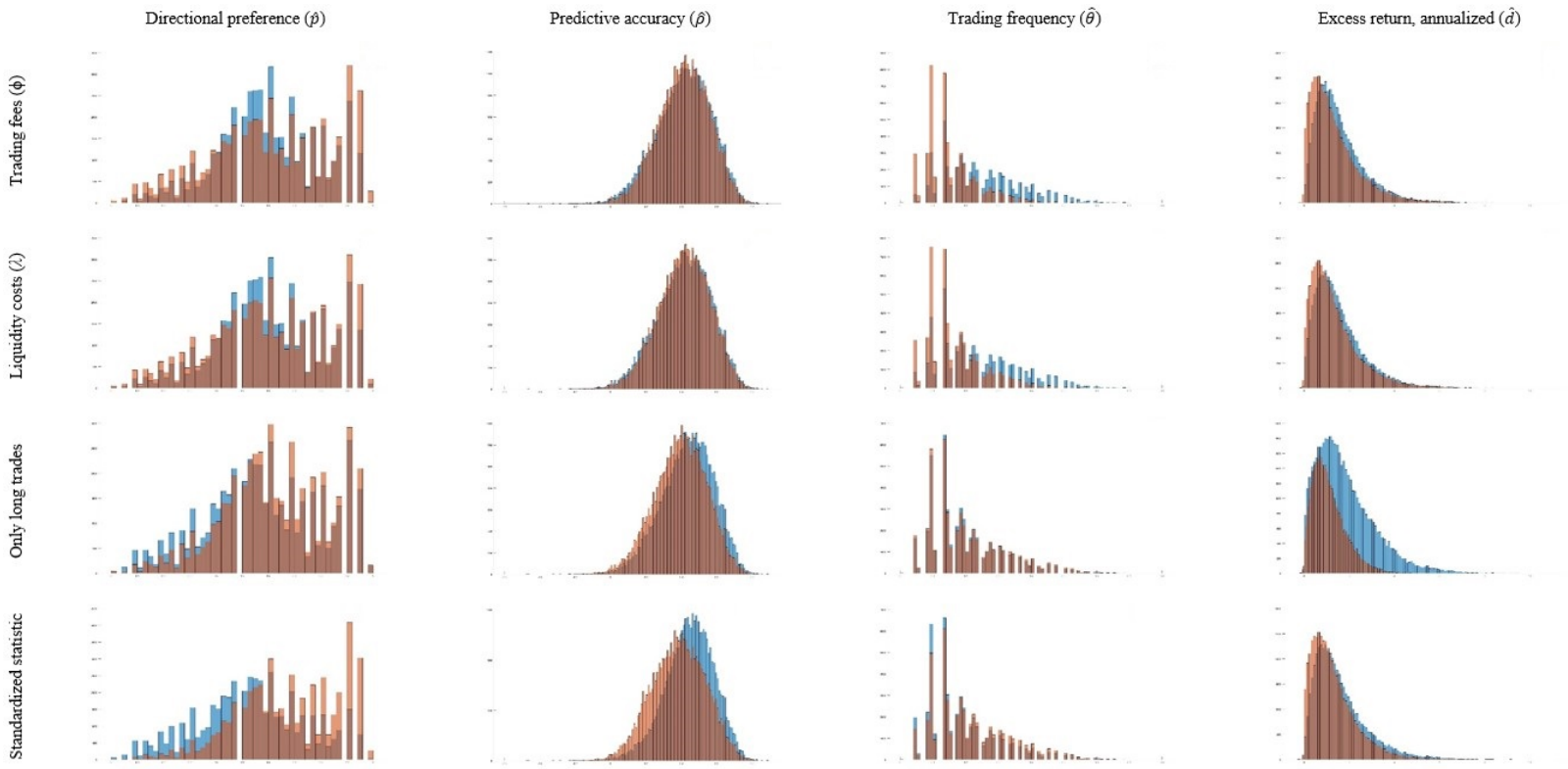

Figure 1: Methodological choices and simulated distributions of lucky trading rule characteristics

Notes: This figure reports how the simulated distributions of lucky trading rule characteristics change with methodological choices (considering trading fees, considering liquidity costs, restricting short trades/only considering long trades, or standardizing the test statistic). The histograms are overlaid, with orange bars denoting the frequency when the choice is true and blue bars denoting the frequency when the choice is false. Each distribution is plotted from 48,000 data points.

Table 1: The influence of methodological choices and sample properties on the characteristics of lucky trading rules

Coefficient	Dependent variable ($\hat{X}$)			
	$\hat{\rho}$	$\hat{\theta}$	$\hat{p}$	$\hat{d}$
α	0.4485***	0.2575***	0.5036***	0.0389***
	[142.19]	[137.80]	[171.15]	[4.10]
$\hat{\beta}_{fee}$	-0.0079***	-0.1404***	-0.0018	-0.1635***
	[-2.81]	[-84.00]	[-0.71]	[-19.27]
$\hat{\beta}_{liq}$	0.0009	-0.0621***	-0.0026	0.0022
	[0.31]	[-37.19]	[-1.02]	[0.25]
$\hat{\beta}_{long}$	-0.0412***	-0.0006	0.0435***	-0.0388***
	[-14.61]	[-0.36]	[16.54]	[-4.58]
$\hat{\beta}_{std}$	-0.0522***	0.0164***	0.1389***	0.0017
	[-18.51]	[9.82]	[52.78]	[0.21]
$\hat{\gamma}$	0.0024**	-0.0184***	0.1245***	-0.9733***
	[2.37]	[-30.69]	[131.64]	[-319.21]
$\hat{\gamma}_{fee}$	-0.0043***	0.0137***	0.0260***	-0.0013
	[-4.74]	[25.50]	[30.77]	[-0.49]
$\hat{\gamma}_{liq}$	-0.0047***	0.0120***	0.0221***	-0.0036
	[-5.24]	[22.40]	[26.17]	[-1.33]
$\hat{\gamma}_{long}$	-0.0003	-0.0077***	-0.0100***	0.4019***
	[-0.35]	[-14.37]	[-11.92]	[147.37]
$\hat{\gamma}_{std}$	0.0017*	-0.0315***	-0.0160***	0.1155***
	[1.95]	[-58.78]	[-18.99]	[42.36]
$\hat{\delta}$	0.0644***	0.0197***	0.0314***	7.8321***
	[5.86]	[3.03]	[3.06]	[236.87]
$\hat{\delta}_{fee}$	0.0248**	0.2093***	0.0091	-0.2251***
	[2.52]	[35.92]	[0.99]	[-7.61]
$\hat{\delta}_{liq}$	-0.0049	-0.0116**	0.0097	-0.6962***
	[-0.50]	[-1.99]	[1.06]	[-23.54]
$\hat{\delta}_{long}$	0.0103	0.0154***	-0.0057	-2.4589***
	[1.05]	[2.65]	[-0.62]	[-83.14]
$\hat{\delta}_{std}$	-0.0385***	0.0101*	-0.0386***	-1.1275***
	[-3.91]	[1.74]	[-4.20]	[-38.12]
Adjusted R^2	0.0630	0.3648	0.5734	0.8468
F-statistic	460.63***	3939.93***	9218.56***	37921.15***

Notes: This table reports the coefficients (β, γ and δ) estimated for four multivariate linear regression models of the form:

$$\hat{X} = \alpha + \gamma\hat{\xi} + \delta\hat{\sigma} + \sum(\beta_D D + \gamma_D D\hat{\xi} + \delta_D D\hat{\sigma}) + \epsilon$$

where $\hat{X}$ is a characteristic of lucky trading rules (either predictive accuracy$-\hat{\rho}$, trading frequency$-\hat{\theta}$, bullish tendency$-\hat{p}$, or (annualized) average excess return$-\hat{d}$); D is a vector of dummy explanatory variables representing possible methodological choices (considering trading fees$-fee$; considering liquidity costs$-liq$; restricting short trades/only considering long trades$-long$; or standardizing the test statistic$-std$) taking the value of 1 if used and 0 otherwise; $\hat{\xi}$ is the sample average return; $\hat{\sigma}$ is the sample standard deviation of returns; α is the intercept; and ϵ is the error term. Each regression is estimated using 96,000 observations, which are obtained by running the simulation exercise described in Section (3.1) on 6,000 data samples with 16 possible methodological choices. T-statistics are reported in square parenthesis, while ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table 2: Methodological choices and false discoveries–all samples

	fee	No	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Both
	liq	No	No	No	No	Yes	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes	Both
	long	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes	Both
	std	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	Both
Number of trading rules	1	0.522	0.755	0.602	0.787	0.385	0.581	0.386	0.615	0.353	0.524	0.336	0.568	0.268	0.387	0.211	0.481	0.418
	2	0.437	0.714	0.533	0.751	0.317	0.535	0.323	0.575	0.294	0.481	0.276	0.522	0.221	0.344	0.170	0.429	0.374
	4	0.361	0.671	0.454	0.706	0.260	0.484	0.263	0.528	0.238	0.435	0.225	0.477	0.177	0.303	0.130	0.378	0.335
	8	0.326	0.636	0.390	0.677	0.233	0.445	0.219	0.492	0.211	0.398	0.186	0.442	0.159	0.275	0.111	0.344	0.305
	16	0.311	0.622	0.368	0.662	0.224	0.429	0.199	0.475	0.203	0.378	0.169	0.424	0.153	0.258	0.100	0.329	0.290
	32	0.300	0.612	0.365	0.656	0.217	0.421	0.195	0.466	0.196	0.369	0.165	0.415	0.149	0.248	0.098	0.322	0.281
	64	0.212	0.594	0.340	0.647	0.158	0.401	0.171	0.454	0.145	0.349	0.145	0.401	0.108	0.231	0.081	0.294	0.268
	128	0.212	0.594	0.340	0.647	0.158	0.401	0.171	0.454	0.145	0.349	0.145	0.401	0.108	0.231	0.081	0.294	0.268
	256	0.212	0.553	0.335	0.603	0.158	0.359	0.168	0.416	0.144	0.306	0.142	0.363	0.107	0.203	0.078	0.274	0.241
	512	0.158	0.363	0.263	0.419	0.114	0.214	0.120	0.268	0.103	0.184	0.099	0.231	0.077	0.116	0.052	0.183	0.142
	1024	0.131	0.308	0.207	0.365	0.087	0.174	0.087	0.227	0.081	0.147	0.070	0.193	0.057	0.089	0.033	0.148	0.116
	2048	0.128	0.307	0.206	0.365	0.086	0.174	0.086	0.227	0.080	0.146	0.070	0.192	0.056	0.088	0.033	0.147	0.116
	4096	0.122	0.294	0.201	0.351	0.083	0.164	0.083	0.216	0.077	0.138	0.068	0.182	0.054	0.083	0.032	0.141	0.112
	8192	0.113	0.211	0.172	0.268	0.077	0.115	0.071	0.157	0.069	0.097	0.059	0.131	0.049	0.058	0.027	0.110	0.081
	16384	0.105	0.203	0.164	0.260	0.070	0.109	0.066	0.151	0.063	0.092	0.054	0.127	0.047	0.053	0.024	0.104	0.076
	32768	0.091	0.130	0.135	0.178	0.063	0.060	0.052	0.090	0.056	0.052	0.043	0.074	0.039	0.028	0.018	0.072	0.039
	65536	0.090	0.127	0.132	0.174	0.061	0.058	0.051	0.087	0.055	0.050	0.042	0.070	0.039	0.027	0.017	0.070	0.038
	131072	0.088	0.123	0.127	0.168	0.059	0.056	0.049	0.083	0.053	0.047	0.041	0.068	0.038	0.025	0.016	0.067	0.035
	262144	0.082	0.114	0.116	0.156	0.055	0.053	0.043	0.076	0.049	0.043	0.036	0.064	0.034	0.022	0.013	0.062	0.033
	524288	0.079	0.105	0.112	0.145	0.053	0.046	0.040	0.070	0.047	0.039	0.035	0.059	0.032	0.019	0.013	0.057	0.028
	688740	0.072	0.090	0.095	0.126	0.050	0.037	0.033	0.058	0.044	0.030	0.029	0.050	0.029	0.015	0.010	0.049	0.024
ExRet Pctl. 90%		3.683	2.936	2.197	2.101	3.458	2.839	1.891	1.790	3.468	2.835	1.896	1.794	3.300	2.774	1.701	1.578	2.630
ExRet Pctl. 95%		4.375	3.463	2.576	2.454	4.163	3.383	2.234	2.122	4.179	3.402	2.234	2.132	3.997	3.410	2.042	1.909	3.247
ExRet Pctl. 99%		5.898	4.752	3.298	3.147	5.653	4.913	2.987	2.844	5.722	4.797	3.009	2.846	5.514	4.884	2.814	2.672	4.772

Notes: This table reports the size of multiple testing procedures when evaluating lucky trading rules on randomly generated data samples. Type I error rates are estimated at a 5% significance level when varying the methodological choices as follows: considering trading fees–fee; considering liquidity costs–liq; restricting short trades/only considering long trades–$long$; or standardizing the test statistic–std. Each cell value in the main body of the table represents the average error rate over 6,000 test results, corresponding to the number of distinct samples that are defined and used; except for the values in the final (*'Both'*) column, which are aggregated from all 96,000 results. The bottom three rows report critical values for the empirical excess return distribution, i.e. the Qth percentile of the distribution, where Q is 90%, 95%, or 99%, estimated via bootstrap simulation with 1,000 replications.

Table 3: Contribution of different methodological choices to data snooping bias-all samples

Average absolute contribution					Average relative contribution				
TTRs	No Fee	No Liq.	No Short	Std.	TTRs	No Fee	No Liq.	No Short	Std.
1	0.188	0.142	0.026	0.204	1	33.2%	26.2%	2.3%	35.7%
2	0.181	0.137	0.030	0.223	2	35.4%	28.1%	2.9%	41.8%
4	0.171	0.131	0.029	0.234	4	37.6%	30.3%	2.7%	48.0%
8	0.162	0.124	0.022	0.234	8	38.6%	31.0%	-0.1%	51.1%
16	0.160	0.121	0.019	0.231	16	39.5%	31.6%	-2.8%	52.1%
32	0.159	0.120	0.021	0.228	32	39.9%	31.8%	-1.7%	52.3%
64	0.153	0.117	0.042	0.251	64	41.3%	33.5%	8.3%	60.3%
128	0.153	0.117	0.042	0.251	128	41.3%	33.5%	8.3%	60.3%
256	0.148	0.112	0.042	0.217	256	42.2%	33.8%	8.5%	56.8%
512	0.109	0.085	0.038	0.124	512	44.7%	36.2%	10.4%	50.3%
1024	0.096	0.075	0.032	0.112	1024	48.0%	39.8%	6.4%	54.3%
2048	0.096	0.075	0.033	0.113	2048	48.1%	39.8%	7.2%	54.7%
4096	0.092	0.072	0.032	0.106	4096	48.1%	39.8%	7.9%	53.9%
8192	0.073	0.057	0.026	0.064	8192	48.5%	39.6%	6.1%	43.1%
16384	0.071	0.056	0.026	0.063	16384	49.0%	40.3%	5.5%	44.3%
32768	0.052	0.042	0.018	0.023	32768	50.2%	41.6%	2.0%	20.2%
65536	0.051	0.041	0.017	0.022	65536	50.4%	41.6%	0.3%	18.8%
131072	0.050	0.040	0.016	0.021	131072	50.8%	42.1%	-0.4%	17.4%
262144	0.047	0.038	0.014	0.020	262144	51.9%	43.4%	-5.2%	18.5%
524288	0.044	0.036	0.014	0.016	524288	52.0%	44.3%	-2.6%	13.0%
688740	0.038	0.032	0.010	0.012	688740	52.8%	45.0%	-11.9%	4.3%
Min	0.038	0.032	0.010	0.012	Min	33.2%	26.2%	-11.9%	4.3%
Max	0.188	0.142	0.042	0.251	Max	52.8%	45.0%	10.4%	60.3%
Average	0.109	0.084	0.026	0.132	Average	44.9%	36.8%	2.6%	40.5%
StDev	0.052	0.039	0.010	0.094	StDev	6.0%	5.6%	5.5%	17.6%

Notes: This table reports the average absolute and relative contribution of four methodological choices to the total data snooping bias. *No Fee* represents tests where no fixed trading fees are deducted from excess returns. *No Liq.* represents tests where no liquidity/price impact costs are deducted from excess returns. *No Short* represents tests where short trades are restricted/only long trades are possible. *Std.* represents tests where the test statistic is standardized. The absolute contribution of a methodological choice to the total bias is estimated as $\bar{E}^* - \bar{E}$, where $\bar{E}^*$ is the average null rejection (error) rate in tests when the choice is true, and $\bar{E}$ is the average null rejection (error) rate in tests when the choice is false. The relative contribution of a methodological choice to the total bias is $\frac{\bar{E}^* - \bar{E}}{\bar{E}^*}$.

Table 4: Methodological choices, data sample properties, and false discoveries

Panel A. Proportion of false discoveries in samples with a positive average return

	fee	No	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Both
	liq	No	No	No	No	Yes	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes	Both
	long	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes	Both
	std	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	Both
Number of trading rules	1	0.163	0.531	0.255	0.584	0.056	0.272	0.052	0.306	0.043	0.220	0.033	0.263	0.015	0.091	0.006	0.188	0.118
	2	0.093	0.464	0.163	0.519	0.029	0.225	0.028	0.258	0.025	0.181	0.018	0.215	0.008	0.068	0.004	0.149	0.086
	4	0.051	0.396	0.086	0.444	0.014	0.175	0.010	0.210	0.011	0.142	0.008	0.170	0.003	0.054	0.001	0.115	0.065
	8	0.031	0.350	0.024	0.398	0.007	0.148	0.003	0.173	0.006	0.116	0.002	0.140	0.002	0.041	0.000	0.093	0.054
	16	0.023	0.330	0.012	0.377	0.005	0.137	0.000	0.160	0.005	0.103	0.001	0.128	0.001	0.035	0.000	0.085	0.048
	32	0.020	0.321	0.011	0.369	0.004	0.133	0.000	0.156	0.004	0.099	0.001	0.123	0.001	0.033	0.000	0.083	0.045
	64	0.000	0.300	0.006	0.357	0.000	0.118	0.000	0.148	0.000	0.086	0.000	0.115	0.000	0.028	0.000	0.075	0.042
	128	0.000	0.300	0.006	0.357	0.000	0.118	0.000	0.148	0.000	0.086	0.000	0.115	0.000	0.028	0.000	0.075	0.042
	256	0.000	0.260	0.005	0.309	0.000	0.099	0.000	0.126	0.000	0.068	0.000	0.095	0.000	0.022	0.000	0.064	0.035
	512	0.000	0.115	0.001	0.146	0.000	0.035	0.000	0.051	0.000	0.027	0.000	0.036	0.000	0.006	0.000	0.027	0.009
	1024	0.000	0.071	0.000	0.103	0.000	0.021	0.000	0.035	0.000	0.015	0.000	0.025	0.000	0.003	0.000	0.017	0.004
	2048	0.000	0.070	0.000	0.103	0.000	0.021	0.000	0.035	0.000	0.015	0.000	0.024	0.000	0.003	0.000	0.017	0.004
	4096	0.000	0.064	0.000	0.094	0.000	0.018	0.000	0.030	0.000	0.012	0.000	0.022	0.000	0.002	0.000	0.015	0.004
	8192	0.000	0.032	0.000	0.053	0.000	0.008	0.000	0.014	0.000	0.005	0.000	0.009	0.000	0.001	0.000	0.008	0.002
	16384	0.000	0.028	0.000	0.049	0.000	0.006	0.000	0.012	0.000	0.004	0.000	0.008	0.000	0.000	0.000	0.007	0.002
	32768	0.000	0.009	0.000	0.014	0.000	0.001	0.000	0.003	0.000	0.001	0.000	0.003	0.000	0.000	0.000	0.002	0.000
	65536	0.000	0.008	0.000	0.014	0.000	0.001	0.000	0.003	0.000	0.001	0.000	0.002	0.000	0.000	0.000	0.002	0.000
	131072	0.000	0.007	0.000	0.012	0.000	0.001	0.000	0.003	0.000	0.001	0.000	0.002	0.000	0.000	0.000	0.002	0.000
	262144	0.000	0.006	0.000	0.010	0.000	0.001	0.000	0.002	0.000	0.001	0.000	0.002	0.000	0.000	0.000	0.001	0.000
	524288	0.000	0.004	0.000	0.009	0.000	0.001	0.000	0.002	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.001	0.000
	688740	0.000	0.003	0.000	0.007	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.001	0.000
ExRet Pctl. 90%		2.346	1.889	1.457	1.393	2.083	1.660	1.159	1.095	2.094	1.645	1.168	1.106	1.883	1.490	0.975	0.895	1.596
ExRet Pctl. 95%		2.695	2.256	1.641	1.598	2.424	2.035	1.346	1.276	2.429	2.044	1.366	1.318	2.214	1.859	1.158	1.075	1.967
ExRet Pctl. 99%		3.284	3.072	2.064	1.994	3.029	2.796	1.742	1.658	3.061	2.859	1.776	1.718	2.880	2.541	1.508	1.429	2.740

Notes: This table reports the size of multiple testing procedures when evaluating lucky trading rules on randomly generated data samples. Type I error rates are estimated at a 5% significance level when varying the methodological choices as follows: considering trading fees–fee; considering liquidity costs–liq; restricting short trades/only considering long trades–$long$; or standardizing the test statistic–std. Each cell value in the main body of the table represents the average error rate over 2,961 test results, corresponding to the number of samples on which average returns are positive; except for the values in the final (*Both*) column, which are aggregated from all 47,376 results. The bottom three rows report critical values for the empirical excess return distribution, i.e. the Qth percentile of the distribution, where Q is 90%, 95%, or 99%, estimated via bootstrap simulation with 1,000 replications when average sample returns are positive.

Panel B. Proportion of false discoveries in samples with a negative average return

	fee	No	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Both
	liq	No	No	No	No	Yes	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes	Both
	long	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes	Both
	std	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	Both
Number of trading rules	1	0.872	0.973	0.940	0.986	0.705	0.882	0.711	0.917	0.655	0.821	0.631	0.865	0.514	0.676	0.411	0.767	0.711
	2	0.773	0.957	0.893	0.978	0.599	0.838	0.611	0.884	0.555	0.774	0.528	0.821	0.427	0.613	0.332	0.702	0.654
	4	0.663	0.939	0.813	0.961	0.500	0.784	0.509	0.838	0.458	0.721	0.437	0.775	0.346	0.545	0.256	0.634	0.597
	8	0.613	0.914	0.746	0.948	0.454	0.735	0.430	0.803	0.412	0.673	0.365	0.735	0.312	0.504	0.219	0.588	0.550
	16	0.591	0.905	0.714	0.940	0.437	0.714	0.394	0.782	0.395	0.647	0.333	0.711	0.300	0.475	0.198	0.566	0.525
	32	0.573	0.895	0.708	0.936	0.424	0.702	0.385	0.769	0.382	0.632	0.324	0.698	0.293	0.458	0.193	0.555	0.510
	64	0.419	0.881	0.665	0.930	0.312	0.678	0.338	0.752	0.285	0.606	0.287	0.679	0.212	0.429	0.160	0.508	0.488
	128	0.419	0.881	0.665	0.930	0.312	0.678	0.338	0.752	0.285	0.606	0.287	0.679	0.212	0.429	0.160	0.508	0.488
	256	0.418	0.838	0.657	0.889	0.312	0.612	0.331	0.699	0.284	0.537	0.280	0.625	0.210	0.380	0.155	0.479	0.442
	512	0.312	0.605	0.519	0.686	0.224	0.388	0.237	0.480	0.204	0.337	0.195	0.420	0.151	0.224	0.103	0.335	0.270
	1024	0.259	0.538	0.408	0.619	0.172	0.323	0.172	0.415	0.159	0.275	0.139	0.356	0.113	0.171	0.065	0.276	0.224
	2048	0.253	0.538	0.406	0.619	0.170	0.323	0.170	0.414	0.158	0.274	0.137	0.355	0.111	0.171	0.064	0.274	0.224
	4096	0.241	0.519	0.396	0.600	0.164	0.306	0.165	0.397	0.152	0.261	0.134	0.338	0.107	0.162	0.063	0.264	0.216
	8192	0.222	0.386	0.340	0.477	0.151	0.219	0.141	0.296	0.135	0.186	0.116	0.249	0.097	0.113	0.054	0.209	0.157
	16384	0.208	0.372	0.323	0.466	0.138	0.208	0.130	0.286	0.124	0.177	0.107	0.242	0.092	0.104	0.047	0.198	0.148
	32768	0.179	0.247	0.266	0.337	0.123	0.117	0.103	0.174	0.110	0.101	0.084	0.143	0.078	0.054	0.035	0.139	0.077
	65536	0.177	0.243	0.260	0.331	0.121	0.114	0.101	0.169	0.109	0.097	0.083	0.136	0.077	0.052	0.034	0.136	0.075
	131072	0.173	0.235	0.251	0.320	0.116	0.109	0.097	0.161	0.105	0.092	0.081	0.133	0.075	0.050	0.031	0.131	0.069
	262144	0.162	0.219	0.229	0.298	0.108	0.103	0.084	0.148	0.096	0.084	0.072	0.125	0.067	0.044	0.026	0.121	0.064
	524288	0.157	0.203	0.221	0.277	0.105	0.090	0.078	0.135	0.092	0.076	0.068	0.115	0.063	0.038	0.025	0.112	0.055
	688740	0.143	0.173	0.188	0.242	0.099	0.073	0.065	0.113	0.086	0.060	0.056	0.098	0.056	0.030	0.020	0.097	0.047
ExRet Pctl. 90%		4.371	3.428	2.567	2.440	4.150	3.367	2.227	2.112	4.166	3.380	2.227	2.124	3.985	3.406	2.039	1.904	3.217
ExRet Pctl. 95%		5.108	3.944	2.946	2.790	4.890	3.962	2.603	2.447	4.912	3.947	2.608	2.470	4.714	4.040	2.415	2.253	3.851
ExRet Pctl. 99%		6.328	5.270	3.614	3.372	6.158	5.492	3.233	3.066	6.155	5.426	3.316	3.095	5.999	5.516	3.076	2.910	5.335

Notes: This table reports the size of multiple testing procedures when evaluating lucky trading rules on randomly generated data samples. Type I error rates are estimated at a 5% significance level when varying the methodological choices as follows: considering trading fees–fee; considering liquidity costs–liq; restricting short trades/only considering long trades–$long$; or standardizing the test statistic–std. Each cell value in the main body of the table represents the average error rate over 3,039 test results, corresponding to the number of samples on which average returns are negative; except for the values in the final (*Both*) column, which are aggregated from all 48,624 results. The bottom three rows report critical values for the empirical excess return distribution, i.e. the Qth percentile of the distribution, where Q is 90%, 95%, or 99%, estimated via bootstrap simulation with 1,000 replications when average sample returns are negative.

Panel C. Contribution of samples with a negative average return to the total data snooping bias

	fee	No	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Both
	liq	No	No	No	No	Yes	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes	Both
	long	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes	Both
	std	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	Both
Number of trading rules	1	68.8	29.7	57.7	25.9	85.4	53.2	86.4	50.3	87.8	58.1	90.1	53.7	94.4	76.6	97.3	60.9	71.8
	2	78.8	35.0	69.3	31.0	91.0	58.0	91.2	55.1	91.6	62.4	93.6	58.7	96.2	80.2	97.6	65.3	77.0
	4	86.0	41.0	81.1	37.1	94.5	63.8	96.3	60.2	95.2	67.5	96.3	64.3	98.1	82.1	99.2	69.5	80.6
	8	90.4	45.0	93.8	41.2	97.0	66.9	98.8	64.8	97.3	71.0	98.9	68.3	98.9	85.2	99.7	72.9	82.3
	16	92.5	46.9	96.7	43.0	97.6	68.1	100.0	66.3	97.7	72.8	99.6	69.8	99.1	86.3	100.0	74.0	83.3
	32	93.3	47.5	96.8	43.7	98.0	68.3	100.0	66.6	98.1	73.2	99.6	70.3	99.3	86.7	100.0	74.3	83.9
	64	100.0	49.5	98.3	44.8	100.0	70.7	100.0	67.5	100.0	75.4	100.0	71.4	100.0	87.9	100.0	74.5	84.4
	128	100.0	49.5	98.3	44.8	100.0	70.7	100.0	67.5	100.0	75.4	100.0	71.4	100.0	87.9	100.0	74.5	84.4
	256	100.0	53.0	98.4	48.7	100.0	72.3	100.0	69.7	100.0	77.7	100.0	73.9	100.0	89.4	100.0	76.8	85.3
	512	100.0	68.4	99.7	65.2	100.0	83.7	100.0	81.0	100.0	85.5	100.0	84.3	100.0	95.1	100.0	85.4	93.3
	1024	100.0	76.9	100.0	71.7	100.0	88.0	100.0	84.5	100.0	89.6	100.0	87.2	100.0	96.2	100.0	88.3	96.2
	2048	100.0	77.1	100.0	71.6	100.0	88.0	100.0	84.7	100.0	89.6	100.0	87.3	100.0	96.2	100.0	88.3	96.2
	4096	100.0	78.2	100.0	73.1	100.0	89.1	100.0	85.9	100.0	91.0	100.0	88.1	100.0	97.6	100.0	89.1	96.4
	8192	100.0	84.8	100.0	80.2	100.0	92.9	100.0	91.0	100.0	94.4	100.0	92.8	100.0	98.8	100.0	92.9	97.5
	16384	100.0	86.0	100.0	81.3	100.0	94.1	100.0	91.7	100.0	95.2	100.0	93.3	100.0	99.4	100.0	93.3	97.3
	32768	100.0	93.0	100.0	92.0	100.0	98.3	100.0	96.6	100.0	98.0	100.0	96.3	100.0	100.0	100.0	97.3	99.1
	65536	100.0	93.9	100.0	92.1	100.0	98.3	100.0	96.9	100.0	98.0	100.0	96.6	100.0	100.0	100.0	97.4	99.1
	131072	100.0	94.2	100.0	92.6	100.0	98.2	100.0	96.7	100.0	98.6	100.0	97.0	100.0	100.0	100.0	97.6	99.0
	262144	100.0	94.9	100.0	93.5	100.0	98.7	100.0	96.9	100.0	98.4	100.0	97.4	100.0	100.0	100.0	97.8	99.0
	524288	100.0	95.8	100.0	93.5	100.0	98.5	100.0	96.6	100.0	99.1	100.0	97.1	100.0	100.0	100.0	97.9	100.0
	688740	100.0	96.2	100.0	94.4	100.0	99.1	100.0	97.1	100.0	100.0	100.0	97.3	100.0	100.0	100.0	98.3	100.0

Notes: This table reports the relative (%) contribution of bearish samples to the total data snooping bias, i.e. the number of false discoveries obtained on samples with negative average returns divided by the number of false discoveries estimated on all samples. Type I error rates are estimated at a 5% significance level when varying the methodological choices as follows: considering trading fees–fee; considering liquidity costs–liq; restricting short trades/only considering long trades–$long$; or standardizing the test statistic–std.

Table 5: False discovery rates by average return and standard deviation of returns deciles: RC and SPA tests, 1-month samples

Panel A. All tests

Average return decile	Standard deviation of returns decile									
	1	2	3	4	5	6	7	8	9	10
1	0.955	0.920	0.863	0.844	0.821	0.737	0.680	0.650	0.693	0.566
2	0.718	0.675	0.572	0.520	0.484	0.410	0.369	0.321	0.273	0.297
3	0.477	0.383	0.355	0.307	0.309	0.262	0.261	0.263	0.236	0.232
4	0.255	0.198	0.232	0.200	0.197	0.206	0.204	0.182	0.214	0.174
5	0.144	0.150	0.145	0.126	0.143	0.156	0.144	0.147	0.176	0.138
6	0.083	0.097	0.100	0.102	0.102	0.112	0.124	0.118	0.125	0.095
7	0.029	0.051	0.061	0.066	0.089	0.071	0.078	0.065	0.075	0.110
8	0.012	0.014	0.038	0.032	0.047	0.044	0.039	0.070	0.058	0.079
9	0.008	0.008	0.009	0.011	0.017	0.027	0.028	0.036	0.055	0.060
10	0.006	0.000	0.001	0.002	0.004	0.006	0.008	0.014	0.017	0.025

Panel B. Only tests that control for all 688,740 TTRs

Average return decile	Standard deviation of returns decile									
	1	2	3	4	5	6	7	8	9	10
1	0.903	0.769	0.584	0.536	0.495	0.369	0.307	0.291	0.329	0.152
2	0.336	0.272	0.165	0.097	0.077	0.043	0.039	0.013	0.017	0.019
3	0.090	0.041	0.023	0.033	0.030	0.006	0.022	0.012	0.002	0.011
4	0.020	0.008	0.012	0.006	0.005	0.005	0.009	0.008	0.017	0.003
5	0.004	0.007	0.007	0.000	0.000	0.006	0.006	0.007	0.012	0.004
6	0.003	0.001	0.005	0.007	0.002	0.001	0.010	0.003	0.002	0.000
7	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.000	0.004
8	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.001	0.002	0.004
9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005
10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Notes: This table reports the proportion of false discoveries estimated for the RC and SPA tests in the simulation exercise for a significance level of 5%, grouped by average returns and standard deviation of returns deciles. The reported values are aggregated from the entire set of $6,000 \times 16 \times 21 = 2,016,000$ test results (Panel A) or from a set of $6,000 \times 16 = 96,000$ results that are obtained when only using the full set of 688,740 TTRs (Panel B). The data is generated using a GBM with no drift ($\mu = 0$) and volatility $\sigma \in \{0.15, 0.20, 0.25, 0.30, 0.35, 0.40\}$. Sample length is 1 month (21 observations on average). Average (annualized) return decile cutoff points are: -1.2348, -0.7626, -0.4758, -0.2405, -0.0200, 0.2070, 0.4640, 0.7679 and 1.2524. Standard deviation of returns decile cutoff points are: 0.1505, 0.1793, 0.2053, 0.2336, 0.2638, 0.2947, 0.3252, 0.3576 and 0.3994.

Table 6: Drivers of false discoveries

Linear Model			Logistic Model		
Coefficient	Single H_0	Multiple H_0	Coefficient	Single H_0	Multiple H_0
α	0.0820***	0.3578***	α	0.0309	-0.7030***
	[49.95]	[97.23]		[0.48]	[-4.52]
β_{fee}	0.1096***	0.3057***	β_{fee}	-1.9905***	-2.2644***
	[74.62]	[92.88]		[-34.34]	[-14.73]
β_{liq}	0.0486***	0.1251***	β_{liq}	-0.3865***	-0.7287***
	[33.08]	[38.01]		[-6.71]	[-5.08]
β_{long}	0.0069***	0.0510***	β_{long}	0.1726***	-0.2684**
	[4.74]	[15.50]		[3.01]	[-1.99]
β_{std}	-0.0704***	-0.0021	β_{std}	0.4458***	-0.3403**
	[-47.95]	[-0.65]		[7.71]	[-2.39]
γ	0.0635***	0.2364***	γ	-3.6208***	-7.0374***
	[120.33]	[199.86]		[-77.59]	[-46.85]
γ_{fee}	0.0258***	0.0077***	γ_{fee}	-0.1274***	-0.1408
	[54.68]	[7.31]		[-3.49]	[-1.45]
γ_{liq}	0.0210***	0.0082***	γ_{liq}	-0.1754***	-0.2155**
	[44.47]	[7.76]		[-4.86]	[-2.27]
γ_{long}	0.0000	0.0006	γ_{long}	0.2278***	1.0572***
	[0.16]	[0.63]		[6.58]	[11.80]
γ_{std}	-0.0410***	-0.0343***	γ_{std}	0.9490***	3.4017***
	[-86.92]	[-32.50]		[24.18]	[26.32]
δ	-0.0160***	0.0109	δ	0.8615***	-37.1855***
	[-2.79]	[0.85]		[3.58]	[-34.98]
δ_{fee}	-0.2185***	-0.5302***	δ_{fee}	0.8652***	-0.0394
	[-42.65]	[-46.21]		[4.02]	[-0.05]
δ_{liq}	-0.0436***	-0.0079	δ_{liq}	-3.9339***	-5.1633***
	[-8.51]	[-0.69]		[-18.16]	[-6.54]
δ_{long}	-0.0191***	-0.0985***	δ_{long}	0.0227	7.9359***
	[-3.74]	[-8.59]		[0.10]	[10.98]
δ_{std}	0.0458***	-0.0421***	δ_{std}	5.5390***	21.4303***
	[8.95]	[-3.67]		[25.31]	[22.75]
Adjusted R^2	0.5685	0.7023	McFadden R^2	0.5120	0.6273
F-statistic	9035.40***	16181.15***	F-statistic	68075.23***	23730.20***

Notes: This table reports the coefficients (β, γ and δ) estimated for the multivariate linear and logistic regressions: $\hat{X} = \alpha + \gamma\hat{\xi} + \delta\hat{\sigma} + \sum(\beta_D D + \gamma_D D\hat{\xi} + \delta_D D\hat{\sigma}) + \epsilon$, and $\mathbb{1}_{\{\hat{X}\leq 0.05\}} = \frac{1}{1+\exp\{-(\hat{X}=\alpha+\gamma\hat{\xi}+\delta\hat{\sigma}+\sum(\beta_D D+\gamma_D D\hat{\xi}+\delta_D D\hat{\sigma})+\epsilon)\}}$, where $\hat{X}$ is the p-value of the RC test (when the statistic is not standardized) or the p-value of the SPA test (when the statistic is standardized); D is a vector of dummy explanatory variables representing possible methodological choices (considering trading fees$-fee$; considering liquidity costs$-liq$; restricting short trades/only considering long trades$-long$; or standardizing the test statistic$-std$) taking the value of 1 if used and 0 otherwise; $\hat{\xi}$ is the sample average return; $\hat{\sigma}$ is the sample standard deviation of returns; α is the intercept; and ϵ is the error term. *Single* H_0 refers to single hypothesis tests that use only the luckiest rule in each sample. *Multiple* H_0 refers to full-scale multiple hypothesis tests that evaluate the luckiest rule vs. the complete set of 688,740 alternatives. Each regression is estimated using 96,000 observations, which are obtained by running the simulation exercise described in Section (3.1) on 6,000 data samples with 16 possible methodological choices. T-statistics are reported in square parenthesis, while ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table 7: False discovery rates by average return and standard deviation of returns deciles: RC test, 1-year samples

Panel A. All tests

Average return decile \ Standard deviation of returns decile	1	2	3	4	5	6	7	8	9	10
1	0.891	0.705	0.723	0.645	0.606	0.547	0.449	0.465	0.408	0.462
2	0.452	0.393	0.278	0.273	0.244	0.194	0.176	0.174	0.156	0.183
3	0.229	0.161	0.153	0.154	0.124	0.116	0.114	0.108	0.086	0.072
4	0.090	0.073	0.082	0.070	0.071	0.068	0.083	0.065	0.065	0.073
5	0.038	0.028	0.053	0.054	0.040	0.058	0.056	0.051	0.055	0.084
6	0.032	0.025	0.022	0.042	0.028	0.036	0.025	0.025	0.040	0.056
7	0.014	0.017	0.015	0.019	0.024	0.034	0.027	0.035	0.043	0.030
8	0.006	0.012	0.009	0.016	0.012	0.022	0.017	0.028	0.034	0.031
9	0.003	0.004	0.008	0.008	0.012	0.013	0.012	0.021	0.018	0.018
10	0.024	0.000	0.006	0.010	0.009	0.009	0.008	0.007	0.011	0.009

Panel B. Only tests that control for all 688,740 TTRs

Average return decile \ Standard deviation of returns decile	1	2	3	4	5	6	7	8	9	10
1	0.429	0.190	0.250	0.150	0.070	0.089	0.039	0.047	0.036	0.099
2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
6	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Notes: This table reports the proportion of false discoveries estimated for the RC test in the simulation exercise for a significance level of 5%, grouped by average returns and standard deviation of returns deciles. The reported values are aggregated from the entire set of $6,000 \times 21 = 126,000$ test results (Panel A) or from a set of $6,000$ results that are obtained when only using the full set of 688,740 TTRs (Panel B). The data is generated using a GBM with no drift ($\mu = 0$) and volatility $\sigma \in \{0.15, 0.20, 0.25, 0.30, 0.35, 0.40\}$. Sample length is 1 year (261 observations on average). Average (annualized) return decile cutoff points are: -0.359, -0.215, -0.130, -0.059, 0.007, 0.072, 0.142, 0.234 and 0.370. Standard deviation of returns decile cutoff points are: 0.152, 0.193, 0.208, 0.247, 0.275, 0.303, 0.336, 0.361 and 0.396.

Table 8: False discovery rates by average return and standard deviation of returns deciles: RC test, 4-year samples

Panel A. All tests

Average return decile	Standard deviation of returns decile									
	1	2	3	4	5	6	7	8	9	10
1	0.815	0.726	0.716	0.597	0.583	0.491	0.475	0.476	0.456	0.465
2	0.491	0.376	0.333	0.277	0.261	0.259	0.188	0.180	0.190	0.158
3	0.222	0.207	0.174	0.119	0.144	0.106	0.138	0.079	0.127	0.099
4	0.106	0.101	0.098	0.090	0.075	0.081	0.070	0.106	0.071	0.057
5	0.054	0.059	0.045	0.057	0.058	0.053	0.065	0.064	0.079	0.082
6	0.026	0.027	0.036	0.037	0.038	0.046	0.040	0.052	0.044	0.050
7	0.019	0.015	0.026	0.028	0.027	0.030	0.034	0.039	0.055	0.036
8	0.012	0.012	0.014	0.017	0.022	0.024	0.021	0.026	0.039	0.034
9	0.011	0.020	0.012	0.009	0.013	0.023	0.017	0.027	0.022	0.026
10	0.000	0.006	0.004	0.007	0.011	0.012	0.013	0.012	0.014	0.014

Panel B. Only tests that control for all 688,740 TTRs

Average return decile	Standard deviation of returns decile									
	1	2	3	4	5	6	7	8	9	10
1	0.125	0.000	0.250	0.121	0.039	0.053	0.045	0.050	0.047	0.038
2	0.000	0.000	0.019	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
6	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Notes: This table reports the proportion of false discoveries estimated for the RC test in the simulation exercise for a significance level of 5%, grouped by average returns and standard deviation of returns deciles. The reported values are aggregated from the entire set of $6,000 \times 21 = 126,000$ test results (Panel A) or from a set of $6,000$ results that are obtained when only using the full set of 688,740 TTRs (Panel B). The data is generated using a GBM with no drift ($\mu = 0$) and volatility $\sigma \in \{0.15, 0.20, 0.25, 0.30, 0.35, 0.40\}$. Sample length is 4 years (1,044 observations on average). Average (annualized) return decile cutoff points are: -0.178, -0.109, -0.064, -0.031, -0.002, 0.032, 0.068, 0.113 and 0.182. Standard deviation of returns decile cutoff points are: 0.151, 0.197, 0.204, 0.249, 0.275, 0.302, 0.344, 0.356 and 0.398.

Table 9: False discovery rates by average return and standard deviation of returns deciles: RC test, 10-year bearish samples

Panel A. All tests

Average return decile \ Standard deviation of returns decile	1	2	3	4	5	6	7	8	9	10
1		1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.943	1.000
3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.918	1.000	0.714
4	1.000	1.000	1.000	1.000	1.000	1.000	0.932	0.786	1.000	0.714
5	1.000	1.000	1.000	1.000	0.857	0.833	0.838	0.714	0.460	0.500
6	1.000	1.000	1.000	0.946	0.738	0.657	0.556	0.476	0.369	0.476
7	1.000	1.000	0.924	0.730	0.649	0.429	0.476	0.476	0.293	0.421
8	1.000	0.868	0.583	0.599	0.531	0.429	0.508	0.357	0.222	0.345
9	0.619	0.690	0.503	0.524	0.413	0.333	0.390	0.219	0.016	0.180
10	0.476	0.214	0.429	0.071	0.095	0.105	0.086	0.028	0.026	0.091

Panel B. Only tests that control for all 688,740 TTRs

Average return decile \ Standard deviation of returns decile	1	2	3	4	5	6	7	8	9	10
1		1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.800	1.000
3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.714	1.000	0.000
4	1.000	1.000	1.000	1.000	1.000	1.000	0.714	0.400	1.000	0.333
5	1.000	1.000	1.000	1.000	0.500	0.500	0.500	0.250	0.000	0.000
6	1.000	1.000	1.000	0.714	0.250	0.100	0.000	0.000	0.000	0.000
7	1.000	1.000	0.600	0.167	0.125	0.000	0.000	0.000	0.000	0.000
8	1.000	0.667	0.000	0.143	0.000	0.000	0.000	0.000	0.000	0.000
9	0.000	0.250	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Notes: This table reports the proportion of false discoveries estimated for the RC test in the simulation exercise for a significance level of 5%, grouped by average returns and standard deviation of returns deciles. The reported values are aggregated from a set of $600 \times 21 = 12,600$ test results (Panel A) or from a set of 600 results that are obtained when only using the full set of 688,740 TTRs (Panel B). The data is generated using a GBM with drift $\mu = -0.20$ and volatility $\sigma \in \{0.15, 0.20, 0.25, 0.30, 0.35, 0.40\}$. Sample length is 10 years (2,610 observations on average). Average (annualized) return decile cutoff points are: -0.327, -0.282, -0.246, -0.219, -0.197, -0.179, -0.154, -0.129 and -0.091. Standard deviation of returns decile cutoff points are: 0.150, 0.196, 0.201, 0.248, 0.275, 0.300, 0.346, 0.353 and 0.397.

Table 10: Proportion of inconsistent tests by average return and standard deviation of returns deciles: pFDR test, 1-year samples

Panel A. All tests

Average return decile	Standard deviation of returns decile									
	1	2	3	4	5	6	7	8	9	10
1	1.000	0.925	0.946	0.420	0.289	0.343	0.467	0.697	0.591	0.655
2	0.816	0.708	0.474	0.323	0.263	0.281	0.332	0.388	0.367	0.374
3	0.415	0.394	0.343	0.331	0.275	0.294	0.369	0.374	0.359	0.342
4	0.369	0.353	0.359	0.299	0.355	0.343	0.384	0.348	0.362	0.349
5	0.272	0.284	0.328	0.298	0.308	0.235	0.340	0.324	0.332	0.338
6	0.241	0.222	0.219	0.287	0.303	0.310	0.316	0.284	0.297	0.320
7	0.122	0.169	0.186	0.225	0.329	0.300	0.287	0.257	0.278	0.237
8	0.062	0.077	0.125	0.224	0.257	0.395	0.276	0.199	0.222	0.253
9	0.018	0.019	0.080	0.246	0.304	0.274	0.347	0.190	0.171	0.166
10	0.000	0.000	0.022	0.257	0.283	0.330	0.157	0.089	0.084	0.074

Panel B. Only tests that control for all 688,740 TTRs

Average return decile	Standard deviation of returns decile									
	1	2	3	4	5	6	7	8	9	10
1	1.000	0.905	0.938	0.300	0.023	0.101	0.289	0.565	0.420	0.504
2	0.833	0.617	0.310	0.129	0.071	0.053	0.048	0.077	0.017	0.078
3	0.120	0.067	0.022	0.130	0.113	0.024	0.103	0.020	0.000	0.000
4	0.009	0.023	0.028	0.055	0.094	0.102	0.086	0.000	0.031	0.048
5	0.000	0.000	0.015	0.067	0.095	0.024	0.043	0.019	0.000	0.000
6	0.000	0.000	0.000	0.073	0.088	0.150	0.094	0.000	0.000	0.020
7	0.000	0.000	0.016	0.066	0.061	0.064	0.066	0.000	0.022	0.000
8	0.000	0.000	0.026	0.043	0.068	0.178	0.017	0.017	0.000	0.022
9	0.000	0.000	0.017	0.088	0.091	0.051	0.117	0.000	0.000	0.000
10	0.000	0.000	0.000	0.071	0.060	0.117	0.015	0.000	0.000	0.000

Notes: This table reports the proportion of *inconsistent pFDR tests* estimated in the simulation exercise for a significance level of 5%, grouped by average returns and standard deviation of returns deciles. *Inconsistent pFDR tests* represent tests in which the proportion of false discoveries exceed the acceptable 5% predefined limit. The reported values are aggregated from a set of $60,00 \times 21 = 126,000$ test results (Panel A) or from a set of $6,000$ results that are obtained when only using the full set of 688,740 TTRs (Panel B). The data is generated using a GBM with no drift ($\mu = 0$) and volatility $\sigma \in \{0.15, 0.20, 0.25, 0.30, 0.35, 0.40\}$. Sample length is 1 year (261 observations on average). Average (annualized) return decile cutoff points are: -0.359, -0.215, -0.130, -0.059, 0.007, 0.072, 0.142, 0.234 and 0.370. Standard deviation of returns decile cutoff points are: 0.152, 0.193, 0.208, 0.247, 0.275, 0.303, 0.336, 0.396 and 0.396.

Table 11: Null rejection rates by average return and standard deviation of returns deciles: cryptocurrency market, RC and SPA tests, 1-month samples

Panel A. All tests

Average return decile	Standard deviation of returns decile									
	1	2	3	4	5	6	7	8	9	10
1	0.798	0.595	0.496	0.405	0.307	0.241	0.200	0.143	0.100	0.067
2	0.530	0.333	0.239	0.203	0.129	0.105	0.076	0.062	0.050	0.032
3	0.266	0.144	0.111	0.093	0.072	0.069	0.050	0.050	0.024	0.025
4	0.116	0.074	0.060	0.057	0.048	0.039	0.039	0.036	0.037	0.028
5	0.061	0.039	0.042	0.040	0.035	0.034	0.029	0.027	0.025	0.027
6	0.156	0.026	0.025	0.021	0.028	0.020	0.020	0.028	0.029	0.015
7	0.016	0.014	0.018	0.017	0.014	0.022	0.016	0.034	0.030	0.034
8	0.007	0.003	0.007	0.007	0.010	0.015	0.016	0.012	0.025	0.012
9	0.005	0.000	0.001	0.004	0.005	0.007	0.011	0.013	0.018	0.017
10	0.000	0.000	0.006	0.001	0.003	0.001	0.004	0.004	0.012	0.015

Panel B. Only tests that control for all 688,740 TTRs

Average return decile	Standard deviation of returns decile									
	1	2	3	4	5	6	7	8	9	10
1	0.333	0.143	0.049	0.024	0.016	0.019	0.011	0.013	0.007	0.011
2	0.037	0.008	0.005	0.008	0.005	0.006	0.009	0.003	0.004	0.002
3	0.011	0.003	0.001	0.005	0.003	0.008	0.000	0.003	0.004	0.000
4	0.009	0.004	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.006
5	0.011	0.002	0.000	0.000	0.003	0.002	0.000	0.000	0.005	0.000
6	0.139	0.005	0.000	0.003	0.002	0.002	0.000	0.005	0.003	0.000
7	0.008	0.003	0.001	0.000	0.004	0.003	0.000	0.004	0.007	0.007
8	0.005	0.000	0.002	0.000	0.000	0.003	0.000	0.000	0.002	0.000
9	0.004	0.000	0.000	0.002	0.000	0.000	0.000	0.001	0.003	0.001
10	0.000	0.000	0.006	0.000	0.002	0.000	0.001	0.000	0.000	0.001

Notes: This table reports the proportion of RC and SPA test null rejections obtained in the empirical exercise conducted on the cryptocurrency market for a significance level of 5%. The results are grouped by average returns and standard deviation of returns deciles. The reported values are aggregated from a set of $32,014 \times 2 \times 21 = 1,344,588$ test results (Panel A) or from a set of $32,014 \times 2 = 64,028$ results that are obtained when only using the full set of 688,740 TTRs (Panel B). Sample length is 1 month (minimum 22 observations; 30 observations on average). Average (annualized) return decile cutoffs are: -8.172, -5.453, -3.638, -2.288, -1.025, 0.079, 1.568, 3.584 and 7.335. Standard deviation of returns decile cutoffs are: 0.893, 1.153, 1.388, 1.644, 1.960, 2.346, 2.884, 3.753 and 5.490.

Table 12: Trading rule overperformance in the cryptocurrency market

Panel A. Simple statistic (RC test)

Cryptocurrency data					Randomly generated data (benchmark)				
TTRs	$\hat{\xi} \in \mathbb{R}$	$\hat{\xi} < 0$	$\hat{\xi} \geq 0$	$\Pr(\hat{\xi} < 0)$	TTRs	$\hat{\xi} \in \mathbb{R}$	$\hat{\xi} < 0$	$\hat{\xi} \geq 0$	$\Pr(\hat{\xi} < 0)$
1	0.102	0.164	0.012	95.0%	1	0.386	0.711	0.052	86.4%
2	0.087	0.141	0.011	94.8%	2	0.323	0.611	0.028	91.2%
4	0.071	0.113	0.010	94.2%	4	0.263	0.509	0.010	96.3%
8	0.055	0.087	0.009	93.0%	8	0.219	0.430	0.003	98.8%
16	0.045	0.070	0.009	92.1%	16	0.199	0.394	0.000	100.0%
32	0.039	0.060	0.008	91.4%	32	0.195	0.385	0.000	100.0%
64	0.029	0.044	0.008	89.0%	64	0.171	0.338	0.000	100.0%
128	0.029	0.044	0.008	89.0%	128	0.171	0.338	0.000	100.0%
256	0.028	0.042	0.008	88.6%	256	0.168	0.331	0.000	100.0%
512	0.016	0.022	0.008	80.7%	512	0.120	0.237	0.000	100.0%
1024	0.008	0.009	0.008	62.5%	1024	0.087	0.172	0.000	100.0%
2048	0.008	0.008	0.008	61.4%	2048	0.086	0.170	0.000	100.0%
4096	0.008	0.008	0.008	59.3%	4096	0.083	0.165	0.000	100.0%
8192	0.006	0.004	0.007	44.6%	8192	0.071	0.141	0.000	100.0%
16384	0.005	0.003	0.007	36.2%	16384	0.066	0.130	0.000	100.0%
32768	0.004	0.002	0.007	23.0%	32768	0.052	0.103	0.000	100.0%
65536	0.004	0.001	0.007	20.5%	65536	0.051	0.101	0.000	100.0%
131072	0.004	0.001	0.007	19.8%	131072	0.049	0.097	0.000	100.0%
262144	0.004	0.001	0.007	15.7%	262144	0.043	0.084	0.000	100.0%
524288	0.004	0.001	0.007	14.9%	524288	0.040	0.078	0.000	100.0%
688740	0.003	0.001	0.007	11.8%	688740	0.033	0.065	0.000	100.0%

Panel B. Standardized statistic (SPA test)

Cryptocurrency data					Randomly generated data (benchmark)				
TTRs	$\hat{\xi} \in \mathbb{R}$	$\hat{\xi} < 0$	$\hat{\xi} \geq 0$	$\Pr(\hat{\xi} < 0)$	TTRs	$\hat{\xi} \in \mathbb{R}$	$\hat{\xi} < 0$	$\hat{\xi} \geq 0$	$\Pr(\hat{\xi} < 0)$
1	0.303	0.452	0.090	87.7%	1	0.615	0.917	0.306	50.3%
2	0.283	0.424	0.082	88.1%	2	0.575	0.884	0.258	55.1%
4	0.256	0.383	0.073	88.1%	4	0.528	0.838	0.210	60.2%
8	0.226	0.340	0.064	88.4%	8	0.492	0.803	0.173	64.8%
16	0.196	0.297	0.052	89.1%	16	0.475	0.782	0.160	66.3%
32	0.170	0.259	0.043	89.5%	32	0.466	0.769	0.156	66.6%
64	0.148	0.226	0.036	90.0%	64	0.454	0.752	0.148	67.5%
128	0.148	0.226	0.036	90.0%	128	0.454	0.752	0.148	67.5%
256	0.130	0.199	0.031	90.3%	256	0.416	0.699	0.126	69.7%
512	0.068	0.103	0.017	89.5%	512	0.268	0.480	0.051	81.0%
1024	0.051	0.078	0.014	88.6%	1024	0.227	0.415	0.035	84.5%
2048	0.051	0.077	0.014	88.7%	2048	0.227	0.414	0.035	84.7%
4096	0.048	0.072	0.013	88.3%	4096	0.216	0.397	0.030	85.9%
8192	0.036	0.053	0.012	86.1%	8192	0.157	0.296	0.014	91.0%
16384	0.031	0.045	0.012	84.2%	16384	0.151	0.286	0.012	91.7%
32768	0.016	0.020	0.011	71.9%	32768	0.090	0.174	0.003	96.6%
65536	0.016	0.019	0.011	71.3%	65536	0.087	0.169	0.003	96.9%
131072	0.015	0.018	0.011	69.3%	131072	0.083	0.161	0.003	96.7%
262144	0.014	0.016	0.011	67.2%	262144	0.076	0.148	0.002	96.9%
524288	0.013	0.014	0.011	65.1%	524288	0.070	0.135	0.002	96.6%
688740	0.012	0.012	0.011	61.3%	688740	0.058	0.113	0.002	97.1%

Note: This table reports the proportion of positive discoveries regarding trading rule overperformance in the cryptocurrency market, i.e. test null rejections at the 5% significance level divided by the total number of tests performed. In total, 32,014 RC and 32,014 SPA tests are performed on data samples containing between 22 and 31 observations, out of which 18,935 are bearish ($\xi \leq 0$) and 13,079 are bullish ($\xi > 0$). $\Pr(\hat{\xi} < 0)$ denotes the proportion of false discoveries obtained in downward trending samples, i.e. the ratio between null rejections obtained on samples exhibiting negative average returns and the total number of null rejections. The benchmark results are obtained from Table 2.

Table 13: Null rejection rates by average return and standard deviation of returns deciles: stock market, RC test

Panel A. 1-year subsamples (231 observations on average)

Average return decile	Standard deviation of returns decile									
	1	2	3	4	5	6	7	8	9	10
1	0.667	0.000	0.133	0.115	0.027	0.025	0.013	0.015	0.036	0.023
2	0.015	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
6	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Panel B. Full samples (3310 observations on average)

Average return decile	Standard deviation of returns decile									
	1	2	3	4	5	6	7	8	9	10
1	0.000	0.000	0.000	0.000	0.000	0.077	0.000	0.079	0.000	0.017
2	0.000	0.000	0.000	0.000	0.050	0.000	0.034	0.000	0.036	0.000
3	0.000	0.069	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.025	0.000	0.000	0.000	0.077	0.000	0.000	0.000
5	0.000	0.056	0.000	0.065	0.056	0.040	0.063	0.000	0.000	0.000
6	0.000	0.057	0.000	0.074	0.037	0.000	0.000	0.000	0.000	0.000
7	0.000	0.036	0.000	0.067	0.000	0.056	0.000	0.000	0.000	0.111
8	0.125	0.000	0.000	0.000	0.056	0.033	0.027	0.000	0.059	0.000
9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.040	0.031	0.071
10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.022	0.000

Notes: This table reports the proportion of RC test null rejections obtained in the empirical exercise conducted on the stock market for a significance level of 5%. The results are grouped by average returns and standard deviation of returns deciles. The values reported in Panel A are aggregated from a set of 34,887 results obtained when testing the full set of 688,740 TTRs on 1-year subsamples (minimum 65 observations; 232 observations on average). In this case, average (annualized) return decile cutoffs for are: -0.536, -0.241, -0.094, 0.013, 0.105, 0.202, 0.309, 0.455 and 0.708; while standard deviation of returns decile cutoffs are: 0.187, 0.226, 0.260, 0.294, 0.330, 0.371, 0.425, 0.502 and 0.638. The values reported in Panel B are aggregated from a set of 2,406 results obtained when testing the full set of 688,740 TTRs on full samples (minimum 65 observations; 3,338 observations on average). In this case, average (annualized) return decile cutoffs for are: -0.131, -0.021, 0.032, 0.064, 0.090, 0.117, 0.150, 0.208 and 0.323; while standard deviation of returns decile cutoffs are: 0.263, 0.305, 0.338, 0.368, 0.404, 0.452, 0.524, 0.654 and 1.038.

Table 14: Trading rule overperformance in the foreign exchange market

Simple statistic (RC test)					Standardized statistic (SPA test)				
TTRs	$\hat{\xi} \in \mathbb{R}$	$\hat{\xi} < 0$	$\hat{\xi} \geq 0$	$\Pr(\hat{\xi} < 0)$	TTRs	$\hat{\xi} \in \mathbb{R}$	$\hat{\xi} < 0$	$\hat{\xi} \geq 0$	$\Pr(\hat{\xi} < 0)$
1	0.476	0.771	0.260	68.4%	1	0.554	0.814	0.365	62.0%
2	0.289	0.600	0.063	87.5%	2	0.398	0.686	0.188	72.7%
4	0.217	0.514	0.000	100.0%	4	0.325	0.629	0.104	81.5%
8	0.193	0.457	0.000	100.0%	8	0.289	0.600	0.063	87.5%
16	0.145	0.343	0.000	100.0%	16	0.277	0.600	0.042	91.3%
32	0.145	0.343	0.000	100.0%	32	0.265	0.571	0.042	90.9%
64	0.145	0.343	0.000	100.0%	64	0.241	0.514	0.042	90.0%
128	0.145	0.343	0.000	100.0%	128	0.241	0.514	0.042	90.0%
256	0.145	0.343	0.000	100.0%	256	0.205	0.429	0.042	88.2%
512	0.120	0.286	0.000	100.0%	512	0.084	0.200	0.000	100.0%
1024	0.084	0.200	0.000	100.0%	1024	0.084	0.200	0.000	100.0%
2048	0.084	0.200	0.000	100.0%	2048	0.084	0.200	0.000	100.0%
4096	0.060	0.143	0.000	100.0%	4096	0.084	0.200	0.000	100.0%
8192	0.060	0.143	0.000	100.0%	8192	0.060	0.143	0.000	100.0%
16384	0.048	0.114	0.000	100.0%	16384	0.048	0.114	0.000	100.0%
32768	0.048	0.114	0.000	100.0%	32768	0.012	0.029	0.000	100.0%
65536	0.048	0.114	0.000	100.0%	65536	0.012	0.029	0.000	100.0%
131072	0.048	0.114	0.000	100.0%	131072	0.012	0.029	0.000	100.0%
262144	0.036	0.086	0.000	100.0%	262144	0.012	0.029	0.000	100.0%
524288	0.036	0.086	0.000	100.0%	524288	0.000	0.000	0.000	
688740	0.024	0.057	0.000	100.0%	688740	0.000	0.000	0.000	

Note: This table reports the proportion of positive discoveries regarding trading rule overperformance in the FX market, i.e. test null rejections at the 5% significance level divided by the total number of tests performed. In total, 21 RC and 21 SPA tests are performed on 83 data samples averaging 973 observations, out of which 35 are bearish ($\xi \leq 0$) and 48 are bullish ($\xi > 0$). $\Pr(\hat{\xi} < 0)$ denotes the proportion of false discoveries obtained in downward trending samples, i.e. the ratio between null rejections obtained on samples exhibiting negative average returns and the total number of null rejections.

Appendix A. Methodological choices, data sample properties, and the characteristics of lucky trading rules—detailed results

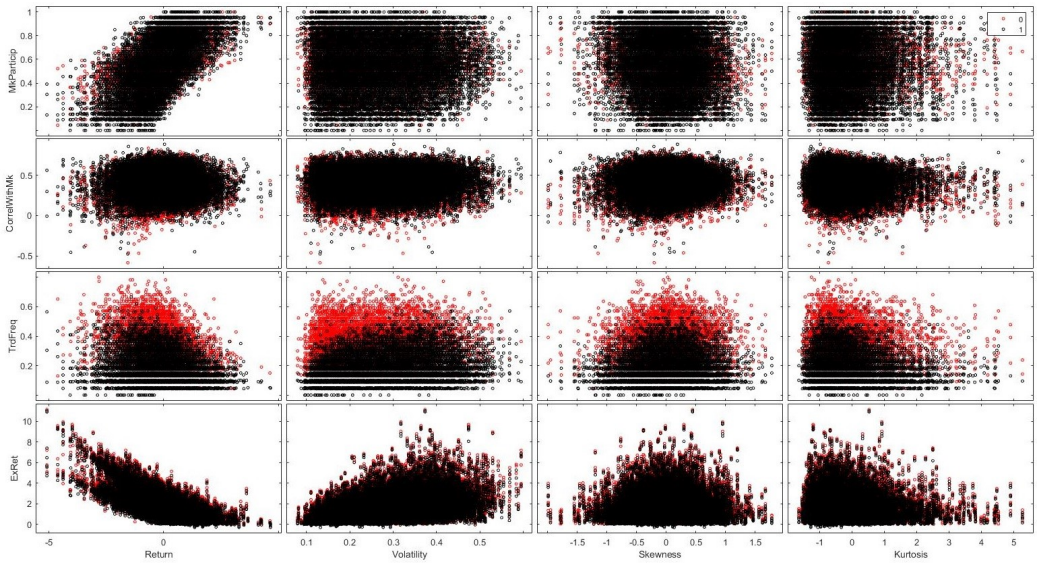

Figure A.1: Data sample properties and the characteristics of lucky trading rules—Grouped by Trading Fee Choice.

Notes: This figure reports how the choice of considering trading fees alters the interaction between lucky trading rule characteristics (either bullish tendency-*MkParticip*; predictive accuracy-*CorrelWithMkt*; trading frequency-*TrdFreq*; or (annualized) average excess return-*ExRet*) and sample properties (average return-*Return*; standard deviation of returns-*Volatility*; *Skewness*; and Excess *Kurtosis*). The dots on the scatter plot are black when the choice is true and red when the choice is false. Each colored group is plotted from 48,000 data points, obtained from running tests on 6,000 data samples with 8 different other combinations of methodological choices.

Main takeaway: The characteristics of lucky trading rules depend on data sample properties. Most notably, the bullish tendency (average excess return) is significantly positively (negatively) influenced by the sample average return. In general, adding trading fees to the loss function reduces the trading frequency of lucky trading rules and slightly decreases their excess return.

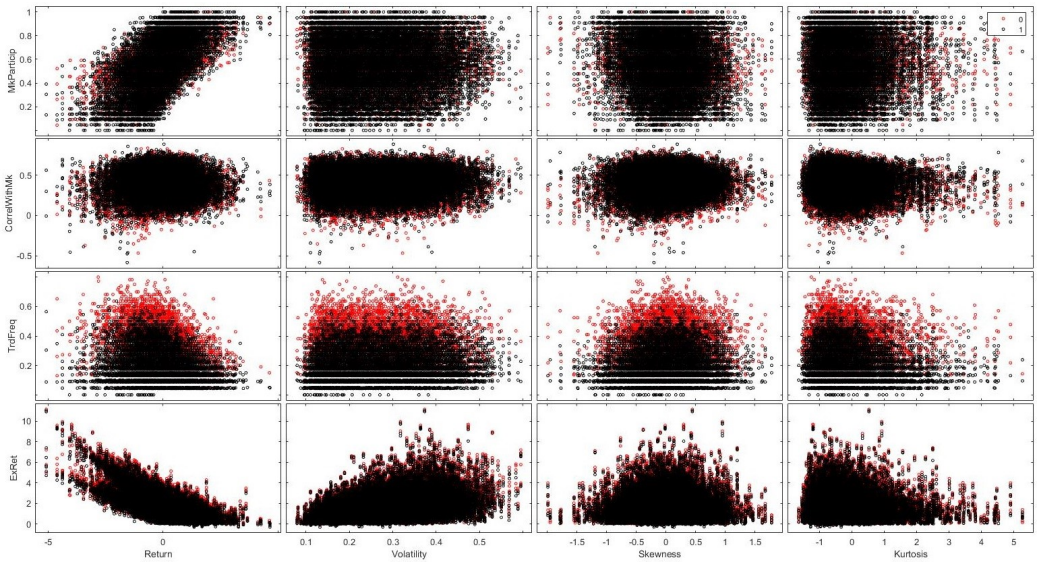

Figure A.2: Data sample properties and the characteristics of lucky trading rules−Grouped by Liquidity Cost Choice.

Notes: This figure reports how the choice of considering liquidity/price impact costs alters the interaction between lucky trading rule characteristics (either bullish tendency-*MkParticip*; predictive accuracy-*CorrelWithMkt*; trading frequency-*TrdFreq*; or (annualized) average excess return-*ExRet*) and sample properties (average return-*Return*; standard deviation of returns-*Volatility*; *Skewness*; and Excess *Kurtosis*). The dots on the scatter plot are black when the choice is true and red when the choice is false. Each colored group is plotted from 48,000 data points, obtained from running tests on 6,000 data samples with 8 different other combinations of methodological choices.

Main takeaway: The characteristics of lucky trading rules depend on data sample properties. Most notably, the bullish tendency (average excess return) is significantly positively (negatively) influenced by the sample average return. In general, adding liquidity costs to the loss function reduces the trading frequency of lucky trading rules and slightly decreases their excess return.

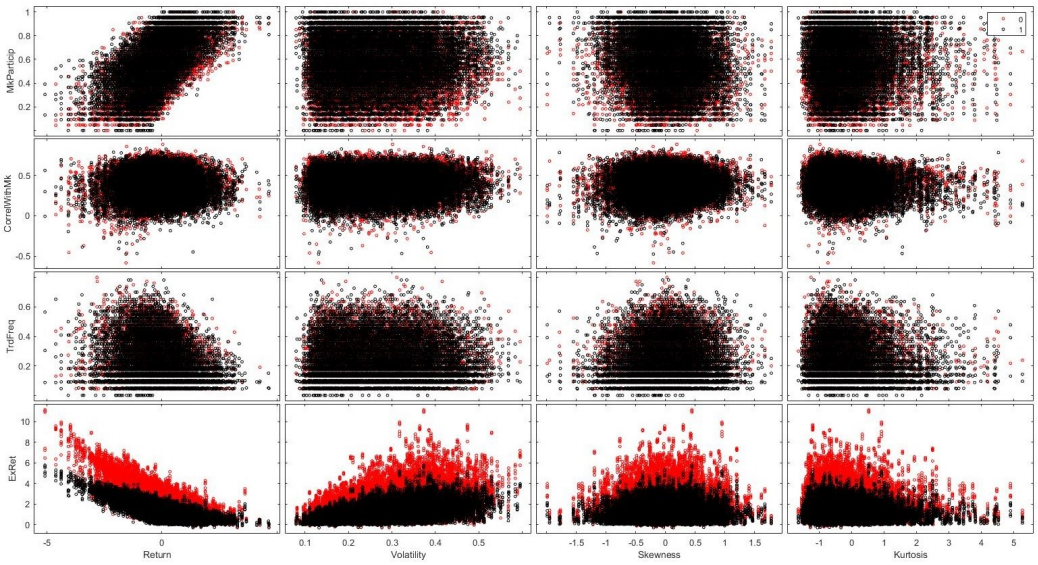

Figure A.3: Data sample properties and the characteristics of lucky trading rules−Grouped by Short Selling/Long Only Choice.

Notes: This figure reports how the choice of restricting short trades (only considering long trades) alters the interaction between lucky trading rule characteristics (either bullish tendency-*MkParticip*; predictive accuracy-*CorrelWithMkt*; trading frequency-*TrdFreq*; or (annualized) average excess return-*ExRet*) and sample properties (average return-*Return*; standard deviation of returns-*Volatility*; *Skewness*; and Excess *Kurtosis*). The dots on the scatter plot are black when the choice is true and red when the choice is false. Each colored group is plotted from 48,000 data points, obtained from running tests on 6,000 data samples with 8 different other combinations of methodological choices.

Main takeaway: The characteristics of lucky trading rules depend on data sample properties. Most notably, the bullish tendency (average excess return) is significantly positively (negatively) influenced by the sample average return. In general, restricting short trades greatly reduces the excess return of lucky trading rules. In particular, it decreases the slope of the relationship between market average returns and trading rule excess returns.

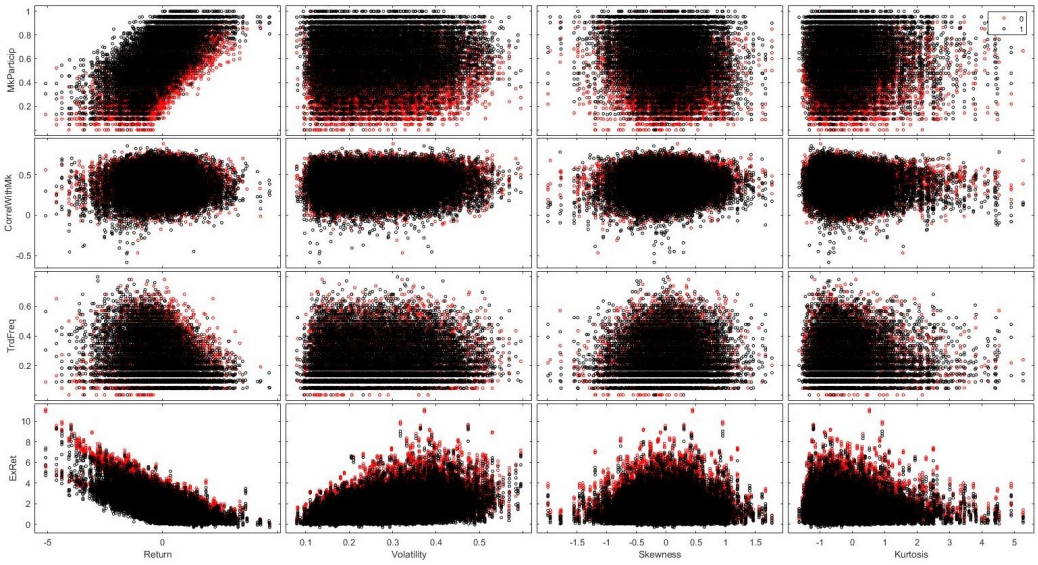

Figure A.4: Data sample properties and the characteristics of lucky trading rules−Grouped by Choice of Test Statistic.

Notes: This figure reports how the choice of standardizing the test statistic alters the interaction between lucky trading rule characteristics (either bullish tendency-*MkParticip*; predictive accuracy-*CorrelWithMkt*; trading frequency-*TrdFreq*; or (annualized) average excess return-*ExRet*) and sample properties (average return-*Return*; standard deviation of returns-*Volatility*; *Skewness*; and Excess *Kurtosis*). The dots on the scatter plot are black when the choice is true and red when the choice is false. Each colored group is plotted from 48,000 data points, obtained from running tests on 6,000 data samples with 8 different other combinations of methodological choices.

Main takeaway: The characteristics of lucky trading rules depend on data sample properties. Most notably, the bullish tendency (average excess return) is significantly positively (negatively) influenced by the sample average return. In general, standardizing the test statistic slightly increases the predictive accuracy of lucky trading rules, but slightly decreases their trading frequency and excess return.