Journal of Financial Economics 106 (2012) 473–491

Contents lists available at SciVerse ScienceDirect

Journal of Financial Economics

journal homepage: www.elsevier.com/locate/jfec

Technical trading revisited: False discoveries, persistence tests, and transaction costs ☆

Pierre Bajgrowicz [a], Olivier Scaillet [b,*]

[a] Université de Genève, and Litasco SA, 40 Bd du Pont d'Arve, 1211 Geneva, Switzerland
[b] Université de Genève and Swiss Finance Institute, 40 Bd du Pont d'Arve, 1211 Geneva, Switzerland

ARTICLE INFO

Article history:
Received 23 November 2010
Received in revised form
5 December 2011
Accepted 6 December 2011
Available online 16 June 2012

JEL classification:
C12
C15
G11
G14

Keywords:
Technical trading
False discovery rate
Persistence
Transaction costs

ABSTRACT

We revisit the apparent historical success of technical trading rules on daily prices of the Dow Jones Industrial Average index from 1897 to 2011, and we use the false discovery rate (FDR) as a new approach to data snooping. The advantage of the FDR over existing methods is that it selects more outperforming rules, which allows diversifying against model uncertainty. Persistence tests show that, even with the more powerful FDR technique, an investor would never have been able to select ex ante the future best-performing rules. Moreover, even in-sample, the performance is completely offset by the introduction of low transaction costs. Overall, our results seriously call into question the economic value of technical trading rules that has been reported for early periods.

1. Introduction

Whether technical trading rules can consistently generate profits, as opposed to just being lucky every now and then, is the subject of an ongoing debate. Practitioners have devoted significant resources to technical trading, which uses past price and volume data to infer future prices. A substantial segment of the investment industry employs indicators that include moving averages, support and resistance levels, and other filter rules. Technical indicators are as ubiquitous on professional information systems as on popular finance websites and online retail brokers. In spite of its popularity among practitioners, academics have long been skeptical about the merits of technical analysis.

☆ We would like to thank the editor and the referee for constructive criticism and numerous suggestions that have lead to substantial improvements over previous versions of the paper. We are grateful to M. Franscini-Scaillet for helping us to get the data on fund structure costs and futures trading costs via her industry contacts. We thank L. Barras, I. Chaieb, M. Dubois, B. Dumas, L. Frésard, E. Ghysels, R. Gibson, R. Gonzalez, P. Hsu, M. Rockinger, A.Timmermann, and A. Treccani, as well as seminar participants at the finance seminar of the University of Athens (2009), the Econometric Society European Meeting (2008), the European Finance Association conference (2008), the Bachelier Finance Society World Congress (2008), the European Financial Management Association annual meeting (2008), the Computational and Financial Econometrics workshop (2008), the Society for Financial Econometrics (SoFiE) inaugural conference (2008), the Swiss Society for Financial Market Research conference (2008), the Meeting of the Swiss Society of Economics and Statistics (2008), the Midwest Finance Association meeting (2008), the Brown Bag Seminar of the University of Zurich (2008), the University of Rome Workshop on Quantitative Finance (2008), the EC^2 conference (2007), the Swiss Doctoral Workshop in Finance (2007), and the Association Française de Finance conference (2007) for helpful comments. We received financial support from the Swiss National Science Foundation through the National Center of Competence in Research: Financial Valuation and Risk Management (NCCR FINRISK).

* Corresponding author. Tel.: +41 22 379 88 16; fax: +41 22 379 81 04.
E-mail addresses: pbajgrowicz@litasco.ch (P. Bajgrowicz), olivier.scaillet@unige.ch (O. Scaillet).

0304-405X/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jfineco.2012.06.001

They argue that it is inconsistent with the theory of market efficiency, which states that all available information must be reflected in security prices. In hopes of resolving this conflict, researchers have undertaken numerous empirical studies of technical trading rules. Some have found results in favor of the ability of trading rules to deliver superior returns, e.g., Neftci (1991), Brock, Lakonishok, and LeBaron (BLL, 1992) Neely, Weller, and Dittmar (1997), Sullivan, Timmermann, and White (STW, 1999) Lo, Mamaysky, and Wang (2000), and Kavajecz and Odders-White (2004). Other studies conclude that trading rules cannot be used to predict future prices. For example, Fama and Blume (1966), Bessembinder and Chan (1998), Allen and Karjalainen (1999), and Ready (2002) show that transaction costs outweigh the predictive power of trading rules. In addition to the impact of transaction costs, researchers have warned against the danger of data snooping which raises the possibility that the reported results are spurious. Menkhoff and Taylor (2007) provide an extensive review of the literature on the use of technical analysis in foreign exchange markets.

In this paper, we revisit the apparent historical success of trading rules during early time periods found in previous studies, including studies reaching an overall negative conclusion such as Ready (2002). In particular we examine the performance of the 7,846 trading rules of STW on daily prices of the Dow Jones Industrial Average (DJIA) index between January 1897 and July 2011. The first contribution is to apply the false discovery rate (FDR) methodology developed by Barras, Scaillet, and Wermers (2010) in the context of mutual funds selection, as a new approach to select outperforming rules while accounting for data snooping. We show that the FDR approach has numerous advantages compared with existing methods. The second contribution is to perform a rigorous analysis of the economic value of the trading rules. We focus on two issues that have been only partly addressed in the literature: the impact of transaction costs and the question of whether investors could have reasonably selected the future outperforming rules without the benefit of foresight. Equipped with the more powerful FDR approach to detect rules with true predictive ability and accounting for transaction costs ex ante, we perform persistence tests in which we measure the out-of-sample performance of the selected rules. We are the first to carry out such a comprehensive persistence analysis of trading rules. Only by combining all these relevant factors can the economic value of the strategies be truly assessed.

To illustrate the problem of data snooping, imagine you put enough monkeys on typewriters and that one of the monkeys writes *The Iliad* in ancient Greek. Because of the sheer size of the sample, you are likely to find a lucky monkey once in a while. Would you bet any money that he is going to write *The Odyssey* next? The same principle applies to trading rules. By looking long enough and hard enough on a given set of data, an investor always finds a trading rule parameterization that works, even if it does not genuinely possess predictive power. For a discussion of the dangers of data snooping, see Lo and MacKinlay (1990), White (2000), and the references therein. Diebold (2006) also warns against the danger of in-sample overfitting. Kosowski, Naik, and Teo (2007) study the impact on detecting hedge fund performance.

In this paper we propose a new methodology to select superior trading rules while accounting for data snooping based on the FDR. We employ the FDR^+ and the FDR^-, developed by Barras, Scaillet, and Wermers (2010). The $FDR^{+/-}$ gives the proportion of false discoveries—rules with no genuine performance, separately among the rules selected as delivering statistically significant positive and negative performance. As we show in a Monte Carlo experiment, the FDR approach has advantages compared with statistical methods used in previous studies, e.g., the bootstrap reality check (BRC) of White (2000) employed by STW, and its stepwise extension by Romano and Wolf (RW, 2005). The BRC indicates only whether the rule that performs best in the sample beats the benchmark, after accounting for data snooping. It provides no information on the other strategies. In practice, investors prefer not to base their investment decision on a single strategy. Though potentially able to detect further outperforming rules, the RW method relies on the conservative familywise error rate (FWER), which results in a lack of power; see Romano, Shaikh, and Wolf (2008b) for a discussion. One further problem with methods derived from the BRC such as the RW method is that they do not select further strategies once they find a rule whose performance is due to luck, even if there remain an important number of true outperforming rules in the population. The FDR approach by tolerating a certain (small) proportion of false discoveries, does not suffer from the problem. We run a Monte Carlo study calibrated to the setting of our empirical work and taking into account the cross-sectional dependence of trading strategies. The Monte Carlo simulations illustrate that situations in which a rule with no genuine predictive power achieves one of the highest performance are common in practice. They also show that the FDR approach greatly improves the chances of detecting all true outperforming rules and behaves well even if the rules are not independent. Using the FDR method, an investor can construct a portfolio of rules on which to base his investment decision and, hence, diversify against model risk.

With the help of our new more powerful rules selection approach, we investigate whether the trading rules can make money. BLL show examples of historical performance and consider them as proof of the usefulness of the trading rules. STW argue that the findings of BLL are not spurious as the best rule passes the BRC data snooping test. However, although it can be the case that we are able to find rules that perform well historically, no indication exists that we can select these rules ex ante. Another important issue not addressed ex ante in BLL and STW is the impact of transaction costs. The rules selected before transaction costs produce very frequent trading signals, and their predictive power is likely to be offset by transaction costs. Previous studies do not treat transaction costs as endogenous to the selection process. Hence, the relevant question is: Could investors reasonably have anticipated which rules would generate performance outweighing transaction costs? To answer this question, we perform persistence tests, adding a transaction cost each time a buy or sell signal is generated. Specifically, we

measure the out-of-sample performance of a portfolio of rules selected using our new FDR approach and updated every month using data from the previous month. Rebalancing the portfolio monthly has the further advantage of being closer to what is done in practice than previous studies. Investors never get the chance to trade over multiple-year periods before being evaluated, and they update their trading rules regularly in an attempt to adapt to the changing economic environment. The persistence analysis is a major contribution of this paper. STW qualify as out-of-sample the results for the period after the original BLL study but, in fact, they always measure performance in-sample. Persistence analysis has been applied to mutual funds, e.g., Carhart (1997). To our knowledge, however, this is the first time this type of persistence tests are performed on technical trading rules. Jacquier and Yao (2002) implement another approach to persistence analysis also inspired by the mutual fund literature. They follow Brown and Goetzmann (1995) and estimate the probability that a trading rule beats the benchmark over consecutive periods. Their study is limited to the ten moving average rules of BLL and finds that the performance is not persistent at horizons shorter than five years.

Our tests show that, even with our new FDR rule detection approach, the reason for choosing the rules with future superior performance is clear only to researchers examining the price data ex post. Contrary to the mutual fund literature, we conclude that there is no hot hands phenomenon. In addition, even the in-sample historical performance is canceled already with the inclusion of low (conservative) transaction costs. Again, it is only by considering all the relevant aspects—performance persistence, transaction costs, and data snooping—together that we can correctly assess the economic value of the strategies. Our study confirms the results of Ready (2002) and Allen and Karjalainen (1999), who also deal with data snooping and rule selection, though in a very different fashion based on a genetic algorithm.

Our analysis indicates that the past period of predictability reported by numerous studies is not really a puzzle. The BLL results should be viewed as a statistical anomaly, discovered ex post by extensive data snooping. In any case, they should not be viewed as an episode of market inefficiency, as the hypothetical predictability could not have been exploited. Although we provide evidence against the usefulness of the simple trading rules of STW to deliver superior returns when applied in a blue-chip investment environment (DJIA index), our results say little about the existence of profitable trading strategies in other markets or using different trade frequencies. The growing number of institutions getting involved in high-frequency trading hints that profitable algorithmic strategies can be found. Our results do, however, indicate that investors should be wary of the common technical indicators present on any investment website or professional information system and advertised as obvious money-making tools.

Section 2 reviews existing methods to account for data snooping and presents the FDR based approach. Section 3 describes the universe of 7,846 technical trading rules, the performance measurement, and the data. Section 4 illustrates the advantage of the FDR approach by applying it in the same framework as STW. Section 5 presents the persistence analysis, while simultaneously accounting for transaction costs. It also investigates the impact of short sale constraints. Section 6 gathers concluding remarks. Appendices contain technical details on the implementation of the FDR approach and results of Monte Carlo experiments showing the advantages of the FDR method. We also review the literature on gauging total transaction costs and their evolution over time, and we provide up-to-date data for current market conditions. An Appendix with supplementary empirical and simulation results as well as files with the data set and programs used in the paper are posted on the Journal of Financial Economics web page.

2. Data snooping measures

In Section 2.1, we review existing data snooping methods. We present our new approach based on the false discovery rate in Section 2.2, before discussing in Section 2.3 how we can use it to construct a portfolio of trading rules.

2.1. Existing data snooping methods

Data snooping is widely recognized to be a significant issue in the finance literature. Standard methods such as the Bonferroni correction—in which individual tests are performed at the α/l level of significance to guarantee that the significance level of the simultaneous test of all l strategies does not exceed α—are too conservative. A first solution exploiting the dependence structure of the individual test statistics is provided by the bootstrap reality check of White (2000). The BRC provides a procedure to test whether the best rule in the sample has genuine predictive power after accounting for the effects of data snooping. Formally, the BRC tests the null hypothesis that the performance of the best technical trading rule is no better than the performance of the benchmark: $H_0: \max_{k=1,\ldots,l}\varphi_k \leq 0$, where φ_k is the performance measure of the kth rule and is equal to zero when rule k does not generate abnormal performance. The BRC is the data snooping measure used in the study of STW. However, it is not able to identify further strategies that generate true performance. In practice, investors prefer to get a confirmation from multiple strategies. A first attempt to tackle this issue is the stepwise multiple testing method of Romano and Wolf (2005). The RW algorithm uses a modified BRC as a first step and can detect further outperforming strategies in subsequent steps. The RW method controls for the familywise error rate, which is defined as the probability of erroneously selecting one or more trading rules as significant, when in reality they are simply lucky. The FWER is a conservative criterion, resulting in a low power to detect superior performance, especially when the universe of rules is large. Our Monte Carlo study shown in Appendix G illustrates the weakness of the RW method (and of the BRC) in terms of power. Hansen (2005) offers some improvements over the BRC.

Being less sensitive to the influence of poor and irrelevant strategies, his method is more powerful. However, like the BRC, Hansen's method addresses only the question of whether the strategy that appears best in the observed data really beats the benchmark. Hsu and Kuan (2005) utilize the test of Hansen to reexamine the profitability of technical analysis and conclude that there are no profitable trading rules in mature markets [i.e., DJIA and Standard & Poor's (S&P) 500]. Hsu, Hsu, and Kuan (2010) introduce a stepwise extension of the method of Hansen. Using their new test, they find that technical rules have predictive power on growth and emerging markets indices, at least until corresponding exchange traded funds (ETF) are introduced.

2.2. False discovery rate

We now present our new approach based on the false discovery rate to select trading rules while accounting for data snooping. The $FDR^{+/-}$ methodology we use has been developed by Barras, Scaillet, and Wermers (2010) in the context of mutual funds selection. However, our paper is the first to propose the FDR as a tool to account for data snooping.

In practice, investors do not consider the signal of one trading rule at a time but typically combine the signals of multiple strategies. A nontrivial fraction of strategies might possess genuine predictive power, and the goal is to identify a large number of them to diversify against model risk. Benjamini and Hochberg (1995) argue that in such a case the control of the FWER is not necessary. Guarding against any single erroneous detection is much too strict and leads to many missed findings. To identify as many outperforming rules as possible without including too many false positives, Benjamini and Hochberg (1995) propose a more tolerant error measure, the FDR. The basic idea is simple. By allowing a certain (small) proportion of false discoveries, the FDR significantly improves the power of detecting the outperforming rules.

The original FDR paper of Benjamini and Hochberg (1995) assumes that the multiple hypotheses (e.g., trading rules) are independent. Some strategies in our sample are only minor variations of themselves, e.g., moving averages with only slightly different parameters, and are therefore highly correlated. Efforts have been made to generalize the FDR methodology under dependence. For example, Benjamini and Yekutieli (2001) show that we can work under certain dependence conditions, such as positive regression dependency. This covers multivariate normal test statistics with positive correlation and multivariate student test statistics. Storey (2003), Storey and Tibshirani (2003), and Storey, Taylor, and Siegmund (2004) show that when the number of tests l is large the FDR approach holds under "weak dependence" of the p-values (or test statistics). In the multiple testing literature, it is natural to think about large l asymptotics, i.e., to have an increasing number of tests; see, e.g., Finner and Roters (2002). When the number l of tests cannot be taken large, Romano, Shaikh, and Wolf (2008a) show that resampling procedures incorporating information about the dependence structure are better able to detect false null hypotheses.

Farcomeni (2007) and Wu (2008) give several examples illustrating that the notion of weak dependence is general enough to cover many problems of practical interest. Weak dependence can loosely be described as any form of dependence whose effect becomes negligible as the number of tests increases to infinity. The more local the dependence, i.e., the faster dependencies disappear for distant p-values, the more likely it is to satisfy the weak dependence criterion. In our empirical study, the trading rules behave dependently in small groups, with each group being essentially independent of the others. For example, a two-day moving average rule with a 0.01 band is highly correlated to a two-day moving average rule with a 0.015 band. However, the performance of a two hundred-day moving average rule is going to be very different, let alone a filter or a support and resistance rule. Such form of dependence is called block dependence and satisfies the weak dependence conditions. Figs. 2 and 3 in Section 4 illustrate the presence of blocks of similar strategies, with each block behaving differently. Hence, we can safely apply the methods we use to estimate the various parameters of the FDR procedure. In addition, the Monte Carlo simulations that we run in Appendix G confirm the good behavior of our FDR method under cross-sectional dependences.

Elaborating on the FDR, Barras, Scaillet, and Wermers (2010) introduce the $FDR^{+/-}$, which allows to estimate separately the proportion of false discoveries among technical rules that perform better or worse than the benchmark. We call a trading rule significantly positive if its abnormal performance is both significant (i.e., $H_{0k}: \varphi_k = 0$ is rejected in favor of the alternative $H_{Ak}: \varphi_k > 0$ or $\varphi_k < 0$, where φ_k is a performance measure for rule k) and positive. Let R^+ denote the number of trading rules selected as significantly positive. Among them, F^+ do not truly generate abnormal performance but have been selected erroneously. The FDR among the rules yielding positive returns, denoted by FDR^+, is defined as the expected value of the proportion of erroneous selections among the rules selected as outperforming. The FDR^+ can be estimated as $\widehat{FDR}^+ = \widehat{F}^+/\widehat{R}^+$, where $\widehat{F}^+$ and $\widehat{R}^+$ are estimators of F^+ and R^+. Similarly, an estimator of the FDR among the rules yielding negative returns, denoted by FDR^-, can be written as $\widehat{FDR}^- = \widehat{F}^-/\widehat{R}^-$. An FDR^+ of 10% means that among the rules selected as outperforming, on average 10% do no generate genuine positive performance. An FDR^+ of 100% shows that no rule is able to deliver positive returns and that the apparent performance is purely due to luck, i.e., data snooping. An FDR^+ of 0% indicates that all selected strategies do genuinely generate positive performance. The FDR approach also allows for estimating the proportions π_A^+ and π_A^- of, respectively, positive and negative trading rules in the population.

In our application, the FDR offers a sensible balance between true positives and erroneous elections. It is much less conservative than the FWER and leads to a significant increase in power. The FDR approach has received much recent attention in the statistics literature; see Abramovich, Benjamini, Donoho, and Johnstone (2006) for applications of the FDR and for an extensive discussion of the advantages of using the FDR over the FWER in the field of multiple testing.

Romano, Shaikh, and Wolf (2008b) review a number of recent proposals to account for multiple tests, and they discuss how these procedures apply to the problem of model selection. In addition to its less conservative nature, the FDR approach is able to detect the outperforming rules, even if the performance of the best rule in the sample is due to luck, contrary to the RW method and the BRC.

In Appendix G, we design a Monte Carlo experiment replicating the environment of our empirical study, in particular the serial and cross-sectional dependencies. Barras, Scaillet, and Wermers (2010) run an extensive Monte Carlo study that illustrates the good statistical properties of the FDR method in a mutual fund performance measurement setting. Their design covers dependent test statistics, where dependencies come from both the factor structure explaining mutual fund returns and some residual cross-sectional correlations in the error terms. We show in our simulations that the proportions of outperforming and underperforming rules are estimated very accurately by the FDR method for correlated test statistics when l is large. This confirms the behavior predicted by asymptotic theory when l goes to infinity. More important, the FDR approach allows for detecting almost all outperforming rules, while keeping the amount of false discoveries at the desired level. The simulations highlight the lack of power of the RW method and the advantage of the new FDR approach. They illustrate that one reason explaining the low power of the RW method is that its algorithm stops once it encounters a lucky rule. They also show that a situation in which a rule with no genuine predictive power achieves one of the highest performance by luck is not uncommon. At worst, the RW method (and the BRC) selects no single rule if the performance of the best rule in the sample is due to luck. This comes from the stepwise nature of the algorithm controlling the conservative FWER criterion.

Another virtue of the FDR approach is its simplicity. Once the p-values corresponding to the individual tests have been calculated, the estimation of the $\text{FDR}^{+/-}$ is straightforward. The FDR approach requires only p-values from a two-sided test. For each rule k, $1 \leq k \leq l$, we test the null hypothesis H_{0k} of no abnormal performance, versus the alternative H_{Ak} of the presence of abnormal performance, positive or negative: $H_{0k} : \varphi_k = 0, H_{Ak} : \varphi_k > 0$ or $\varphi_k < 0$. The single parameter to be estimated is the proportion π_0 of rules in the population satisfying the null hypothesis $\varphi = 0$. We obtain the individual p-values using the same resampling technique as STW. All relevant estimation procedures to get $\widehat{\text{FDR}}^{+}$, $\widehat{\text{FDR}}^{-}$, $\widehat{\pi}_A^{+}$, and $\widehat{\pi}_A^{-}$, as well as the stationary bootstrap used to obtain the individual p-values, are detailed in the Appendices. We also describe how to determine the standard deviation of the estimators for π_0, π_A^{+}, and π_A^{-} under dependent p-values. These new results extend the asymptotic properties provided by Barras, Scaillet, and Wermers (2010) for independent p-values when l goes to infinity.

2.3. FDR portfolio in practice

Our criterion to construct a portfolio of trading rules sets $\widehat{\text{FDR}}^{+}$ equal to 10%. Just as when choosing the significance level of a statistical test, the choice of the FDR level defines the balance between wrongly including underperforming trading rules and leaving out truly outperforming ones. Our experiments with real data and in our Monte Carlo study indicate that a target of 10% achieves a good trade-off. Results are qualitatively stable for values ranging from 5% to 20%. Another approach useful when we do not know which FDR level to choose is to first fix the rejection region, before computing the corresponding proportion of false discoveries. For example, we select strategies generating positive performance and having a p-value inferior to the threshold $\gamma = 0.01$ in a first step. Then, we compute the resulting $\widehat{\text{FDR}}^{+}$.

We use the algorithm described in the Appendix to pick the corresponding trading rules. We denote the resulting portfolio the 10%-FDR^{+} portfolio. Ninety percent of the rules included in the portfolio possess genuine predictive power. After pooling the signals of the selected rules with equal weight, we invest a proportion of the wealth corresponding to the neutral signals in the risk-free rate and go long or short the market with the remaining money. For example, imagine the 10%-FDR^{+} portfolio contains 60 rules, of which 40 generate a buy signal, 10 generate a neutral signal, and 10 generate a sell signal. After pooling, we obtain 30 buy signals and 20 neutral signals. Hence, we invest 60% of the wealth in the index and the remaining 40% in a savings account. Our portfolio approach is equivalent to averaging the forecasts of the selected rules with equal weights and no prior. Setting more weight on the better rules has an effect very similar to reducing the FDR target level to keep fewer rules. Such a forecast combinations approach that diversifies against model uncertainty is discussed in Elliott and Timmermann (2008).

In theory, we could construct a universe containing all the possible combinations of trading rules and use the BRC to select the best candidate. However, this approach is not feasible in practice as there are $2^{7846}-1$ possible rules combinations, a number with more than two thousand digits. Our FDR portfolio methodology allows us to circumvent this computational hurdle.

3. Trading rules, data, and performance measurement

We describe the universe of trading rules and the data in Section 3.1, and explain how we measure performance in Section 3.2.

3.1. Universe of trading rules and data

When applied to a series of past prices, a trading rule indicates whether a long position (buy), a neutral position (out of the market), or a short position (sell) should be taken in the next time period. To examine whether the apparent success reported in BLL and STW is spurious, i.e., the result of extensive tweaking of the parameters of popular rules, we need to specify a universe of trading rules from which investors could have drawn their strategies. To allow for comparison, we stick to the universe of STW, which consists of l=7,846 rules divided into the following five categories. The technical indicators corresponding to these strategies are very common in practice.

They are available on professional information systems and advertised on popular finance websites.

Filter rules: An investor following a filter rule buys and sells a stock if its price movement reverses direction by a sufficiently high amount. Moves less than a certain percentage in either direction are ignored. The filter rule is supposed to permit investors to participate in a security's major price trends without being misled by small fluctuations.

Moving averages: Investors frequently use moving averages to discover trends in stock prices. For example, in an uptrend, long commitments are retained as long as the price remains above the moving average.

Support and resistance rules: Support and resistance is the concept in technical trading that the movement of the price of a security tends to stop and reverse at certain predetermined price levels. The idea is that the price is more likely to bounce off a support level rather than break through it. However, once the price has passed this level, it is likely to continue dropping until it finds another support level. A resistance level is the opposite of a support level.

Channel breakouts: A price channel is a pair of parallel trend lines that form a trending chart pattern for a security. When the price passes through a trend line, the trend is broken and the breakout generates a buy or sell signal.

On-balance volume averages (*OBV*): The total volume for a given day is assigned a positive or negative value depending on the close being higher or lower than the previous day, and it is added to the OBV of the previous day. The OBV is generally used to confirm price moves. The intuition is that volume is higher on days when the price move is in the dominant direction. Therefore, technical traders consider greater volume on rising prices bullish. Conversely, greater volume on falling prices is considered bearish.

We refer to STW for the exact parameterizations of the trading rules. Apart from the support and resistance rules, which can be considered as contrarian strategies, the other categories of rules are momentum or trend-following strategies. As in STW, we apply the nearly 8,000 trading rules to daily closing prices on the DJIA index. STW consider the sample from January 1897 to December 1996, divided into five subperiods. We add one period for the new data between January 1997 and July 2011. STW also run the strategies on the one hundred-year period from the inception of the DJIA index. Results for this latter sample should be viewed with caution, as market conditions have evolved dramatically in the last one hundred years. Furthermore, managers never get to trade for one hundred years before their performance is evaluated. It can be argued that, in the early periods, it was impossible to trade stock indices frequently without incurring significant transaction costs. With the introduction of exchange-traded funds, e.g., the Diamonds Trust, which tracks the DJIA index, and index futures, it is realistic to assume that investors apply technical rules directly to a stock market index.

3.2. Performance measurement

Each rule k, $1 \le k \le l$, generates an investment signal $s_{k,t-1}$ for each prediction period t, $L \le t \le T$. $s_{k,t-1}$ equals 1 for a long position, 0 for a neutral position, and -1 for a short position. An alternative that leads to the same conclusions on the performance of trading rules is to translate a buy signal into borrowing money at the risk-free rate and doubling the investment in the stock index, a neutral signal into simply holding the index, and a sell signal into exiting the market.

For each rule, we compute a test statistic φ_k, which measures the performance of the rule relative to a benchmark. The statistic is defined in such a way that $\varphi_k = 0$ under the null hypothesis that rule k does not generate abnormal performance relative to the benchmark. Following STW, our benchmark is to be out of the market and to earn the risk-free rate, which corresponds to testing whether the trading rules are able to generate absolute returns. Alternatively, we could compare the performance of the trading rules with a buy-and-hold strategy that is fully invested in the index over the entire sample. A further possibility is to compare the returns of the trading rules with an average of the average returns on the index and on bonds over the period, weighted by the fraction of days the strategy is invested in, respectively, the index and bonds. This allows to test if the trading rule chooses relatively better days to be invested in the index; see Ready (2002).

In their study, STW use two simple performance criteria: the mean return and the Sharpe ratio. We focus our analysis on the Sharpe ratio, which measures the average excess return per unit of total risk. We compute the return in excess of the risk-free rate. This implies that trading rules earn the risk-free rate on days when a neutral signal is generated. We use the same risk-free rate as STW, i.e., the daily federal funds rate after July 1954. We are grateful to A. Timmermann for providing us the DJIA index and risk-free rate series for early periods.

Let y_t be the (arithmetic) period t return on the price series on which the strategies are applied. As in STW, we denote by $f^e_{k,t} = \mathbb{1}_{\{s_{k,t-1} \neq 0\}}(s_{k,t-1} y_t - r_{f,t})$ the period t excess return of rule k, where $r_{f,t}$ is the risk-free rate, and $\mathbb{1}_{\{s_{k,t-1} \neq 0\}} = 1$ if a buy or sell signal is generated and 0 if the signal is neutral. The mean excess return can be written as $\overline{f}^e_k = (1/N)\sum_{t=L}^{T} f^e_{k,t+1}$, and the standard deviation as $\sigma^e_k = \sqrt{(1/(N-1))\sum_{t=L}^{T}(f^e_{k,t+1} - \overline{f}^e_k)^2}$, where $N = T-L+1$ is the number of prediction periods. Then, the test statistic for the Sharpe ratio is simply $\varphi_k = \mathrm{SR}_k = \overline{f}^e_k / \sigma^e_k$. Results for the mean return, which measures the absolute performance, are qualitatively similar, and we do not report them here. They are available in the online Appendix.

As in the available literature, the Sharpe ratios we use for appraising the performance of the trading rules are unconditional, i.e., they involve unconditional risk estimates. The trading rules generate a signal to be either in the market and get market volatility, or out of the market and get a volatility close to zero. The unconditional Sharpe ratio favors rules that are more often out of the market as their denominator is automatically deflated. Taking into account such volatility patterns would likely alter the classification of trading rules. To our knowledge,

this issue is not addressed in the literature. Tackling this issue is not trivial, as the expected value of the conditional Sharpe ratios is not equivalent to the unconditional Sharpe ratio because of Jensen's inequality. This point is less of an issue when assessing the performance of a portfolio of trading rules, which results in being at least partially invested in the index most of the time, as it is the case in our study.

Although it has become a standard in the literature and in the industry, measuring performance with the Sharpe ratio has many drawbacks. The Sharpe ratio does not take into account higher moments and recent studies have shown that incorporating skewness and kurtosis into the portfolio decision causes major changes in the optimal portfolio; see Jondeau and Rockinger (2006) and the references therein. One possible performance measure that can take into account more elaborate utility functions is the certainty equivalent of wealth (CE). The certainty equivalent is that amount of wealth such that the investor is indifferent between receiving it for sure at the horizon and having his current wealth today and the opportunity to invest it up to the horizon; see Brennan, Schwartz, and Lagnado (1997) and Blanchet-Scalliet, Diop, Gibson, Talay, and Tanr (2007). The downside with such an approach is that it requires making an assumption on the stochastic process underlying the index.

4. Long-term in-sample performance

For each of the sample periods, the columns on the right-hand side of Table 1 display the in-sample performance of the best rule in the sample and the corresponding BRC p-value, as reported in STW. The columns on the left-hand side present the performance and size of the portfolio obtained using the RW method to control the FWER at the 5% level. Based on such in-sample evidence discovered ex post, BLL and STW conclude that technical rules can be used to generate profits. These results have no economic value and are merely a test of predictability. They do not take into account transaction costs, and a high historical performance is no indication that an investor could have selected the future best-performing rules in advance. Moreover, in practice, investor performance is evaluated over much shorter periods. Investors update their strategies more frequently, in an attempt to adapt to the changing economic environment. The long-run averages presented in STW tend to mask substantial variability in the rules performance within each period.

Table 1
Performance indicators from existing studies under the Sharpe ratio criterion and with no transaction costs.

This table presents long-term performance results of rules chosen according to the Sharpe ratio criterion, across the different sample periods. The table reports the annualized Sharpe ratio and size of the portfolio obtained with the method of Romano and Wolf (RW, 2005), the annualized Sharpe ratio and corresponding bootstrap reality check (BRC) p-value of the best rule in the sample, and the annualized Sharpe ratio of the buy-and-hold strategy for the Dow Jones Industrial Average index (DJIA). If the portfolio size is zero, the Sharpe ratio is not reported (–).

Sample period	RW portfolio		Best rule		DJIA
	Sharpe ratio	Portfolio size	Sharpe ratio	BRC p-value	Sharpe ratio
1: 1897–1914	1.24	45	1.18	0.00	−0.12
2: 1915–1938	–	0	0.73	0.11	0.06
3: 1939–1962	1.49	62	2.34	0.00	0.41
4: 1962–1986	1.52	15	1.45	0.00	−0.16
5: 1987–1996	–	0	0.84	0.93	0.66
6: 1997–2011	–	0	0.48	1.00	0.12
1897–1996	0.70	88	0.82	0.00	0.12

For the same sample periods, Fig. 1 shows the proportions of outperforming (π_A^+), null (π_0), and underperforming (π_A^-) strategies, estimated using the FDR approach. Results in Fig. 1 are subject to the same criticisms as those in Table 1. However, they illustrate the advantage of our new methodology. For example, in subperiod 2 (1915–1938), the BRC p-value indicates that the performance of the best rule in the sample is not significant after accounting for data snooping. As a consequence, the RW portfolio is empty. However, the FDR analysis reveals that more than 20% of the rules deliver true performance. This example highlights a major problem of the RW method, which is not able to select further rules with genuine performance as soon as it encounters a rule whose performance is due to luck. As illustrated in our Monte Carlo experiments (Appendix G), the event that a rule without true predictive power delivers one of the highest returns by luck is not unlikely.

Fig. 1 indicates that until the 1960s an important proportion of the rules exhibited a significant predictive power. However, predictive power does not imply profitability. An important proportion of the rules that perform best before transaction costs use very short windows of data, generate very frequent trading signals, and, hence, are likely to generate substantial transaction costs. As reported in, e.g., STW or Ready (2002), the rules did poorly in more recent periods. The presence of true underperforming rules before transaction costs, e.g., subperiods 4 (1962–1986), 5 (1987–1996), and 6 (1997–2011), looks counterintuitive at first sight. One might wish to reverse the corresponding signals of when to go in or out of the market. However, most of the systematically negative performance stems from subtracting the risk-free rate from alternating returns of very small magnitude. The trend of poor performance is confirmed by the new data now available for subperiod 6 (1997–2011). As discussed in Ready (2002), one explanation for this drop in performance is that the positive returns of the earlier periods is a statistical anomaly, discovered ex post by extensive data snooping. Another explanation is that an episode of relative market inefficiency did exist, but the predictability was discovered only during more recent periods and became possible to exploit only with lower transaction costs and increased liquidity. Another possible factor is that investors have become more sophisticated and have traded away these opportunities. Friedman (1996) shows

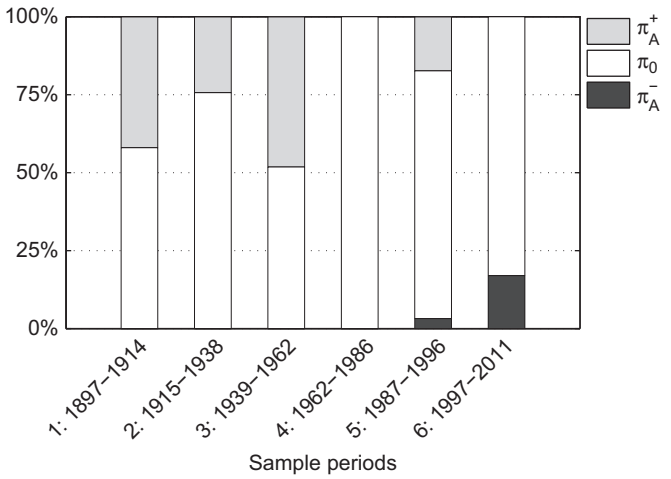

Fig. 1. Proportions of outperforming (π_A^+), null (π_0), and underperforming (π_A^-) rules, under the Sharpe ratio criterion and with no transaction costs. This figure displays estimates of π_A^+, π_0, and π_A^-, across the different sample periods.

Table 2
Transaction costs (TC) such that the long-term in-sample performance disappears under the Sharpe ratio criterion.

This table presents one-way transaction costs in basis points (bps) such that $\widehat{\pi}_A^+$ becomes zero, across the different sample periods. It also displays the corresponding $\widehat{\pi}_A^-$. An asterisk (*) indicates that π_A^- is estimated with zero transaction costs. If $\widehat{\pi}_A^+$ is already zero before TC, TC are not reported (–).

Sample period	TC such that $\widehat{\pi}_A^+ = 0$	Corresponding $\widehat{\pi}_A^-$
1: 1897–1914	16 bps	0%
2: 1915–1938	35 bps	15%
3: 1939–1962	70 bps	36%
4: 1962–1986	–	49%*
5: 1987–1996	–	13%*
6: 1997–2011	–	26%*

that aggregate institutional ownership increases from less than 10% in 1950 to over 50% in 1994. Gompers and Metrick (2001) find that large institutional investors nearly double their share of the stock market from 1980 to 1996. Their increased number does not mean that institutional investors are more sophisticated.

Before we tackle the issue of ex ante rules selection in the next section, we show that the in-sample predictability could not have been turned into profits as the generated returns are not sufficient to outweigh transaction costs. Table 2 reports the minimum transactions costs such that our FDR method does not detect outperforming rules any more. In the first three subperiods (1897–1962), proportional one-way transaction costs as low as 16, 35, and 70 basis points (bps) are sufficient to bring $\widehat{\pi}_A^+$ to zero. Transaction costs are difficult to measure precisely and have declined over time. Nevertheless, studies presented in Appendix H indicate that one-way transaction costs below 50 bps can be considered as conservative starting in the early 1990s and that the costs were significantly higher before the sharp decline triggered by the deregulation of commissions in 1975. Ready (2002) uses one-way transaction costs of 13 bps for the period from 1962 to 1999. Allen and Karjalainen (1999) consider three different one-way transaction costs: 10 bps, 25 bps, and 50 bps, for the period from 1928 to 1995.

The rules selected before transaction costs produce many trading signals, and their performance is canceled once we take into account the costs. In the three most recent sample periods (1962–2011), we do not detect any positive performance ($\widehat{\pi}_A^+ = 0$) already under zero transaction costs. The effective transaction costs depend on a number of factors including the type of trading strategy. For example, a short-term contrarian trading rule will, almost by definition, have a lower price impact than a trend-following strategy. Among the five categories in the STW universe, only the support and resistance rules are contrarian strategies. The other types correspond to momentum or trend-following strategies. As a robustness check, we perform our computations with transaction costs 20% lower for support and resistance rules. The impact on the above results is marginal.

Table 3 illustrates that, once we include transaction costs, the successful rules trade on longer-term price movements. Even if transaction costs have been declining over time, for the sake of comparison we use the same low (i.e., conservative) value of 12.5 basis points across all sample periods. For example, during sample period 3 (1939–1962), if we omit transaction costs, the best rule in the sample uses a window of only two days of data. When transaction costs are taken into account, the best rule needs 250 days, or 12 months of data.

The detailed analysis of the impact of transaction costs, continued in Section 5, is an important contribution of our paper. Although previous studies including BLL and STW call for careful consideration of transaction costs, none provides a satisfactory analysis while simultaneously accounting for data snooping. STW partly address the issue by using price data on the S&P 500 index futures. When trading futures contracts, transaction costs are easy to control, and it is not difficult to take a short position (see Appendix H). However, futures contracts started trading only in 1984, thus limiting the interest of this approach in our one hundred-year sample.

Table 3
Best in-sample technical trading rules under the Sharpe ratio criterion.

This table reports the historically best-performing trading rule chosen with respect to the Sharpe ratio criterion, in each sample period, and for either zero or 12.5 basis points (bps) one-way transaction costs. An asterisk (*) indicates that, according to the bootstrap reality check, the performance of the rule remains significant after accounting for data snooping.

Sample period	Costs	Best trading rule
1: 1897–1914	Zero	5-Day channel rule, 0.02 width, 5-day holding period, 0.005 band*
	12.5 bps	20-Day channel rule, 0.1 width, 5-day holding period
2: 1915–1938	Zero	5-Day moving average, 0.001 band
	12.5 bps	25-Day support & resistance, 5-day delay, 50-day holding period
3: 1939–1962	Zero	2-Day moving average, 0.001 band*
	12.5 bps	75- and 250-day on-balance volume, 0.01 band
4: 1962–1986	Zero	2-Day moving average, 0.001 band*
	12.5 bps	20-Day channel rule, 0.03 width, 10-day holding period, 0.01 band
5: 1987–1996	Zero	50-Day support and resistance, 0.01 band
	12.5 bps	40- and 75-Day on-balance volume, 0.03 band
6: 1997–2011	Zero	Filter rule, 0.01 position initiation, 25-day holding period
	12.5 bps	75- and 250-day on-balance volume, 0.01 band

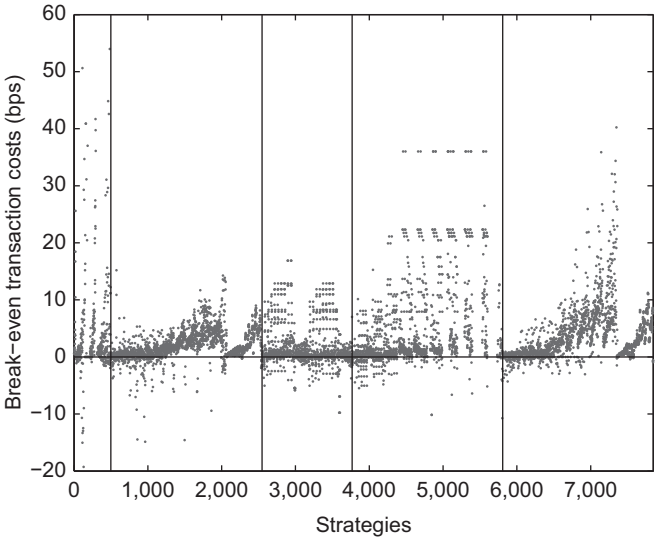

Fig. 2. Break-even transaction costs for the 7,846 trading rules in sample period 3 (1939–1962). Values are given in basis points (bps). Negative values correspond to a negative performance (divided by the number of transactions). The vertical lines separate the different categories of trading rules, which are displayed in the following order: filter rules, moving averages, support and resistance rules, channel breakouts, and on-balance volume averages.

Some studies, e.g., STW and Bessembinder and Chan (1998), compute a break-even transaction cost, which corresponds to the level of transaction costs that exactly offsets the profits from using a given technical trading rule. We also examine break-even costs for each strategy, to see if there is any variation over time or across types of strategies. We do not report the detailed results but Figs. 2 and 3 provide an example of the differences across the various blocks of trading rules and between period 3 (1939–1962) and period 4 (1962–1986). In period 3, 75% of the rules deliver positive in-sample performance before costs, and transaction costs below 25 basis points are sufficient to prevent the vast majority from breaking even. In period 4 the proportion of rules with positive in-sample performance before costs drops to 44%, and most of these rules require costs below 10 basis points to break even. Individual break-even transaction costs are informative. However, it is difficult to use break-even costs in a rules selection process because they are computed ex post, once the trading rules have already been selected. It does not make sense to first select a portfolio of trading rules using the RW method or our FDR approach and then compute the portfolio break-even costs. Trading rules that survive the inclusion of transaction costs are often not among those that perform best before costs. Transaction costs must be treated as endogenous and not exogenous to the selection process. The results of Table 2 and of Section 5 do not suffer from this exogeneity problem. They are obtained by treating transaction costs as endogenous to the selection process, i.e., by increasing transaction costs until the FDR approach is not able to detect any positive performance. They can be viewed as break-even transaction costs computed ex ante. Our approach of computing ex ante break-even costs removes the need to set a level of costs in advance that would be dependent on many factors, such as volume (see Appendix H for a discussion of the price impact as a function of the traded volume).

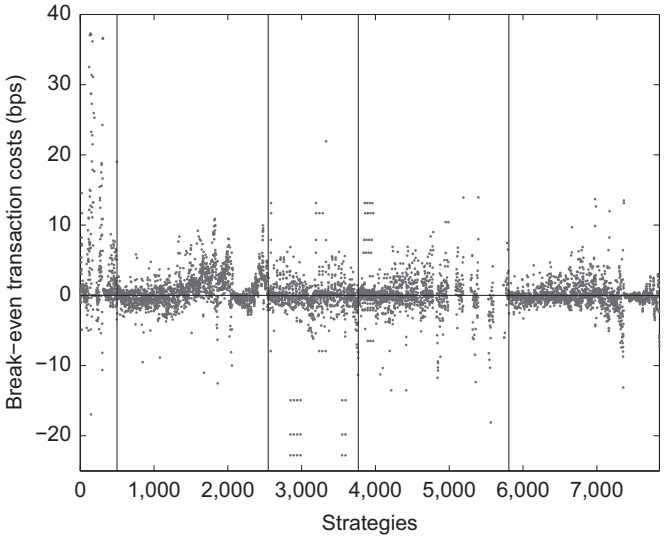

Fig. 3. Break-even transaction costs for the 7,846 trading rules in sample period 4 (1962–1986). Values are given in basis points (bps). Negative values correspond to a negative performance (divided by the number of transactions). The vertical lines separate the different categories of trading rules, which are displayed in the following order: filter rules, moving averages, support and resistance rules, channel breakouts, and on-balance volume averages.

5. Persistence analysis

The question addressed in this section is simple but essential to evaluate the economic value of the trading rules: Could investors reasonably have anticipated which rules would generate superior returns after transaction costs? It is important to ask what information could have been used to select the outperforming rules. If the answer is that the prediction could have been made based on an analysis of investment flows, fiscal policy, or market psychology, then price data alone are not sufficient to reject the assertion. Considering that the trading rules we investigate are purely based on the price action, however, it makes sense to test if the future outperforming rules could have been selected using only past price data. We do so by performing a persistence analysis of the trading rules. Every month, we construct a portfolio of rules using price data of the previous month. We then measure the out-of-sample performance of the selected rules over the following month. It is important to note that to rebalance the portfolio, we use only information that would have been readily available to an investor. Such a persistence analysis of the performance of a large number of trading rules has never been carried out in the literature. STW qualify as out-of-sample the results for the period after the sample of the original BLL study. However, and despite the term, STW always measure the performance in-sample. In our persistence test, rules are selected ex ante and evaluated genuinely out-of-sample. Rebalancing the portfolio monthly as opposed to evaluating the rules over multiple-year periods as is done in all previous studies also allows for selecting different rules depending on the changing economic environment. Compared with existing studies, such setting is much closer to what is done in practice, where investors are evaluated over relatively short time horizons. The one-month period is chosen to correspond to the typical length of a trend in financial markets; see Jegadeesh (1990) and Huang, Liu, Rhee, and Zhang (2010), who find a reversal in stock returns after one month. Results are similar when the rules are updated every six months and are reported in the online Appendix.

Table 4 reports results of the persistence analysis under zero transaction costs, for the same sample periods as previously. It shows the out-of-sample performance for the different rules selection criteria we use, i.e., the 10%-FDR$^+$ portfolio, the RW portfolio, the 50 best-performing rules, and the best rule in the sample. It also displays the median size and the in-sample performance of the monthly rebalanced portfolios. As explained in Section 2.3, we pool the signals of the rules in the portfolio, which results in getting long or short the index with a proportion of the wealth and investing the remaining money at the risk-free rate. The in-sample performance when we update the rules monthly is significantly higher than what we can achieve if we have to keep the same rules over multiple-year periods. However, the out-of-sample performance is negative in most cases throughout the recent periods. Even equipped with the more powerful FDR method, investors could not have reasonably anticipated which rules would generate positive returns, and this even in the unrealistic case of zero transaction costs. Hence, there is no hot hands phenomenon. Other signs show that the reason for choosing the outperforming rules is clear only to researchers examining the price data ex post. The number of selected rules varies greatly from month to month. A study of the portfolio turnover shows that, on average, less than 5% of the rules remain in the portfolio after the first rebalancing. After two rebalancings the portfolio consists of almost exclusively new rules. Although it can be the case that we are able to find ex post technical rules with apparent predictive power, our persistence tests indicate that it is not possible to select these rules ex ante. Our results show that the

Table 4
Performance persistence analysis under the Sharpe ratio criterion and with no transaction costs.

This table displays the in-sample (IS) and the out-of-sample (OOS) annualized Sharpe ratio of trading rules selected according to the following criteria and updated monthly across the different sample periods: the 10%-FDR$^+$ portfolio, the Romano and Wolf (RW, 2005) portfolio, the 50 best rules in-sample, and the best rule in-sample. The table also reports the median size of the false discovery rate (FDR) and RW portfolios across the different rebalancings. If the RW portfolio is always empty, IS performance and OOS performance are not reported (–).

Sample period	FDR portfolio			RW portfolio			50 best rules		Best rule	
	IS	OOS	Median size	IS	OOS	Median size	IS	OOS	IS	OOS
1: 1897–1914	3.41	0.47	14	1.31	0.51	0	5.79	0.50	6.34	0.03
2: 1915–1938	4.62	0.01	13	0.90	0.17	0	5.39	−0.03	5.98	0.09
3: 1939–1962	4.77	0.55	15	1.85	0.09	0	5.78	0.43	6.70	0.12
4: 1962–1986	5.34	−0.31	13	1.36	0.14	0	6.17	−0.18	6.95	−0.59
5: 1987–1996	4.52	−0.34	12	–	–	–	5.44	−0.37	6.07	0.08
6: 1997–2011	4.55	−0.74	12	0.78	0.07	0	5.22	−0.51	5.97	−0.27

Table 5
Transaction costs (TC) such that out-of-sample (OOS) performance disappears under the Sharpe ratio criterion.

This table reports the level of one-way transaction costs in basis points (bps) for which the OOS performance of different portfolios of trading rules becomes zero. As in Table 4, the rules are selected according to the following criteria and updated monthly across the different sample periods: the 10%-FDR$^+$ portfolio, the Romano and Wolf (RW, 2005) portfolio, the 50 best rules in-sample, and the best rule in-sample. The table also displays the median size of the false discovery rate (FDR) and RW portfolios across the different rebalancings. If OOS performance is already zero before TC, TC are not reported (–).

Sample period	FDR portfolio		RW portfolio		50 best rules	Best rule
	TC such that OOS performance=0	Median portfolio size	TC such that OOS performance=0	Median portfolio size	TC such that OOS performance=0	TC such that OOS performance=0
1: 1897–1914	30–35 bps	1	50–55 bps	0	2-0-25 bps	0–5 bps
2: 1915–1938	0–5 bps	12	30–35 bps	0	–	5–10 bps
3: 1939–1962	0–5 bps	7	5–10 bps	0	15–20 bps	0–5 bps
4: 1962–1986	–	–	25–30 bps	0	–	–
5: 1987–1996	–	–	–	–	–	5–10 bps
6: 1997–2011	–	–	20–25 bps	0	–	–

examples of in-sample predictability reported ex post by BLL and STW have no economic value.

Table 4 also displays the advantages of using the FDR approach. The FDR approach efficiently avoids the lucky rules with no genuine performance, as illustrated by the higher out-of-sample returns of the 10%-FDR$^+$ portfolio compared with the performance of the portfolio of the 50 previously best-performing rules. The median size of the different portfolios shows the power advantage of the FDR method, when the RW portfolio is empty most of the times. As explained above and illustrated in the Monte Carlo study (Appendix G), the lack of power of the RW method (and simultaneously of the BRC) comes from the very conservative criteria underlying that method, which prevents it from selecting further rules as soon as it encounters one whose performance is due to luck.

The results of Table 4 show that the performance of trading rules is not persistent and that knowing which rules are going to perform best can be clear only to a person observing the returns ex post. Not to leave any argument in favor of trading rules, we now show that even the smallest transaction costs are sufficient to erase the apparent positive out-of-sample performance still remaining in the early sample periods. Table 5 reports the minimum level of transaction costs so that the out-of-sample performance disappears. As before, we treat the transaction costs as endogenous to the selection process, and we can view the reported levels as break-even transaction costs computed ex ante. Even during the early periods, one-way transaction costs of less than 5–35 basis points suffice to offset any out-of-sample performance. As pointed out in Appendix H, transaction costs were significantly higher in the prevailing periods. Hence, even if the in-sample performance looks attractive, the persistence analysis shows that an investor cannot realistically select the future outperforming rules.

A further source of friction arises from the lending fees when taking a short position. The 10%-FDR$^+$ portfolio results in short positions in more than 20% of the days. Table 6 displays the minimum level of short-selling costs that make the out-of-sample performance disappear. In the first three sample periods (1897–1962), with yearly lending fees only between 5 and 20 basis points (see Appendix H for studies on lending fees in the equity loan market), we are not able to select rules with positive out-of-sample performance, and this while keeping one-way transaction costs at zero. In later periods, out-of-sample performance is already negative before costs.

We have just shown that it is impossible to select the future best-performing rules by looking solely on their past performance. As a robustness check, we test whether other variables can help to predict which trading rules are going to outperform in the future. For example, we test whether certain trading rules perform better within a

Table 6
Lending fees such that the out-of-sample (OOS) performance disappears under the Sharpe ratio criterion.

This table reports the level of yearly lending fees in basis points (bps) such that the OOS performance of the 10%-FDR$^+$ portfolio rebalanced monthly across the different sample periods disappears. It also displays the corresponding median portfolio size across the different rebalancings. If OOS performance is already zero before lending fees, lending fees are not reported (–).

Sample period	Lending fees such that OOS performance = 0	Median portfolio size
1: 1897–1914	10–15 bps	10
2: 1915–1938	0–5 bps	13
3: 1939–1962	15–20 bps	9
4: 1962–1986	–	–
5: 1987–1996	–	–
6: 1997–2011	–	–

particular economic environment, using business cycle data from the National Bureau of Economic Research. Our analysis shows that even knowing the state of the business cycle ex ante would not help an investor selecting the future outperforming rules. We also investigate the predictability of the trading rules conditional on the market environment. For some subperiods, the FDR$^+$ portfolio has a return profile similar to a straddle on the index, i.e., the selected rules perform only when the DJIA index exhibits strong negative or positive returns. Such a pattern has been observed for hedge funds; see Fung and Hsieh (1997). However, the relation is present only in a few sample periods.

6. Conclusion

Previous studies, e.g., BLL and STW, have reported examples of technical trading rules generating superior returns, at least during early time periods. Based on such results observed ex post, they have concluded that trading rules were useful to deliver profits. In our paper, we reassess this apparent historical success.

First, we propose a new approach to select outperforming rules while accounting for data snooping based on the false discovery rate. The FDR method is designed to control false positives, i.e., conclusions that something is statistically significant when it is entirely random. Our Monte Carlo simulations calibrated to our empirical study and taking into account serial and cross-sectional dependencies confirm that the FDR approach is more powerful and better suited than statistical methods used in previous studies. It allows for selecting more rules and diversifying against model uncertainty. Methods derived from the BRC are by construction unable to select further rules once they find a rule whose performance is due to luck. As our simulations illustrate, it is very likely that a rule with no real predictive power achieves by luck a performance better than the majority of the true outperforming rules.

Second, we test whether the trading rules can be used to make money. Because the strategies selected by BLL and STW generate frequent trading signals, return forecastability might not imply superior returns once transaction costs are considered. Another important question is how an investor could have selected the rules able to deliver future returns outweighing transaction costs, without the benefit of hindsight. We address these issues by performing persistence tests of the performance of rules selected with the FDR approach and adding a transaction cost each time a buy or sell signal is generated. Our results show that, in reality, an investor could not have extracted economic value from the simple trading rules of STW in the liquid investment environment (blue-chip index) we consider, even in early sample periods. Even with the help of the more powerful FDR approach, we are not able to select rules whose performance is persistent and not canceled by transaction costs. The rules in the STW universe originate from the Dow Theory of the late 19th century; see Brown, Goetzmann, and Kumar (1998). They are nowadays displayed interactively on popular finance websites and quoted routinely by analysts. They are part of standard packages provided by online brokerage houses and data or news vendors. They can be considered as publicly available, and, in that sense, our results are in favor of the weak efficient-market hypothesis.

However, our results say little about the existence of profitable trading strategies in other markets, using different frequencies or more sophisticated rules. The recent growing number of institutions getting involved in high-frequency trading hints that profitable algorithmic strategies can be found. The same remark applies to the success of statistical arbitrage trading strategies used by several proprietary trading desks and hedge funds in the 1980s and 1990s; see Gatev, Goetzmann, and Rouwenhorst (2006). Our results indicate only that investors should be wary of the common technical indicators present on any investment website or professional information system and advertised as obvious money-making tools.

Appendices

We summarize here results from Barras, Scaillet, and Wermers (2010) and STW, and we present the results of our Monte Carlo experiment showing the better ability of the FDR approach to select the outperforming rules. We also review the literature on transaction costs and short-selling constraints and provide guidelines about which level can be regarded as low (conservative).

Appendix A. Stationary bootstrap

For each trading rule, we test the null hypothesis of no abnormal performance. To obtain the individual p-values, we follow STW and apply the stationary bootstrap of Politis and Romano (1994). This resampling technique is chosen due to the weak correlation in the daily returns. We describe the algorithm that generates a resampled time series of returns. The notation corresponds to that of the text and of STW. Let $\{f_t, t = L, \ldots, T\}$ denote the original series of returns. For $b = 1, \ldots, B$, with $q \in [0, 1]$ a smoothing parameter, the bootstrapped series of returns $\{f_t^b, t = L, \ldots, T\}$ are obtained as follows.

1. Set $t = L$. Draw the index $\theta(t)$ at random, independently and uniformly from $\{L, \ldots, T\}$. Set $f_t^b = f_{\theta(t)}$.

2. Set $t=t+1$. If $t>T$, stop. Otherwise, draw a random variable U from the standard uniform distribution.
   (a) If $U<q$, draw $\theta(t)$ at random, independently and uniformly from $\{L,\ldots,T\}$.
   (b) If $U\geq q$, set $\theta(t)=\theta(t-1)+1$. If $\theta(t)>T$, set $\theta(t)=L$. Set $f_t^b=f_{\theta(t)}$.
3. Repeat step 2.

The stationary bootstrap resamples blocks of varying length from the original data. The average block length equals $1/q$. The parameter q has to be chosen according to the dependence exhibited by the data. We follow STW, who set the average block length to 10 (i.e., $q=0.1$). STW show that the results are robust to the choice of q.

For each simulated series of return, we compute the corresponding performance measure φ^b, $b=1,\ldots,B$. The p-value is obtained by comparing the original performance φ with the quantiles of $\varphi^b-\varphi$, $b=1,\ldots,B$. We set $B=1,000$ for the number of bootstrap iterations.

Appendix B. Estimation of the FDR$^+$ and the FDR$^-$

Suppose that we test the null hypothesis of no abnormal performance for each trading rule and obtain the l corresponding p-values. We call a trading rule significant (i.e., reject the null hypothesis) when its p-value is less than or equal to some threshold γ. Because the null hypothesis we test is two-sided with equal tail significance $\gamma/2$ (see Section 2.2), the false discoveries are spread evenly between outperforming and underperforming trading rules. Based on that observation and following Storey (2003), Barras, Scaillet, and Wermers (2010) propose the following estimators for the FDR separately among the rules yielding positive and negative performance:

$$\widehat{\mathrm{FDR}}^{+}(\gamma)=\frac{\widehat{F}^{+}}{\widehat{R}^{+}}=\frac{\frac{1}{2}\widehat{\pi}_0 l\gamma}{\#\{p_k\leq\gamma,\varphi_k>0;\ k=1,\ldots,l\}} \tag{1}$$

and

$$\widehat{\mathrm{FDR}}^{-}(\gamma)=\frac{\widehat{F}^{-}}{\widehat{R}^{-}}=\frac{\frac{1}{2}\widehat{\pi}_0 l\gamma}{\#\{p_k\leq\gamma,\varphi_k<0;\ k=1,\ldots,l\}}. \tag{2}$$

$\widehat{\pi}_0$ is an estimate of $\pi_0\equiv l_0/l$, the proportion of rules in the population generating no abnormal performance. Hence, measuring the FDR$^{+/-}$ boils down to the estimation of π_0, which we describe in Appendix C.

Appendix C. Estimation of π_0

To estimate π_0, Storey (2002) proposes a method exploiting the fact that, for a two-sided test, null p-values are uniformly distributed over $[0,1]$, whereas p-values of alternative models tend to be close to zero. Fig. A1 shows the histogram density of p-values corresponding to our $l=7,846$ trading rules. We see that, beyond 0.6, the histogram looks fairly flat, which indicates that there are mostly null p-values in this region. The height of this flat portion gives a conservative estimate of the overall proportion of null p-values:

$$\widehat{\pi}_0(\lambda)=\frac{\#\{p_k>\lambda;\ k=1,\ldots,l\}}{l(1-\lambda)}, \tag{3}$$

which involves the tuning parameter λ. It is possible to automate the selection of λ. However, as $\widehat{\pi}_0$ is not sensitive to the choice of λ when the number of rules is high, we set $\lambda=0.6$ by visually examining the histograms. The automated method described in Storey (2002) produces almost identical estimates of π_0.

Appendix D. Estimation of π_A^+ and π_A^-

Appendix C shows how to estimate π_0, from which we can deduce $\pi_A=1-\pi_0$, the proportion of rules with abnormal (i.e., nonzero) performance in the population. It is useful to split π_A into the proportions of rules with

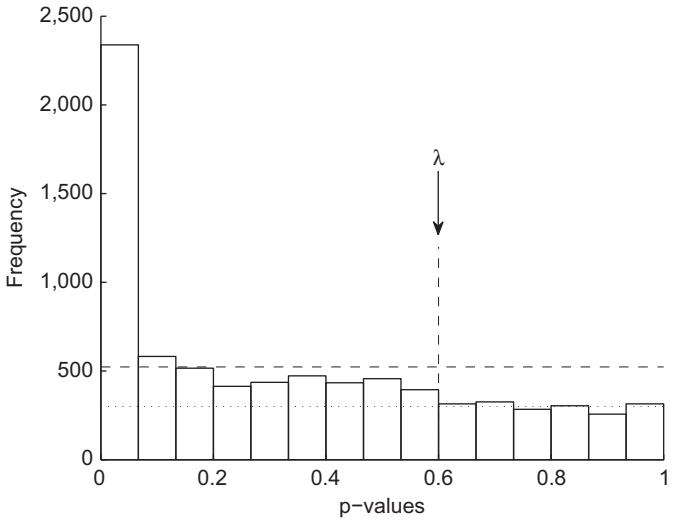

Fig. A1. Density histogram of the 7,846 p-values [sample period 3 (1939–1962), Sharpe ratio criterion]. The dashed line is the density histogram to be expected for the p-values in [0,1] if all rules were truly null (i.e., did not generate abnormal performance). Beyond $\lambda=0.6$, the histogram looks fairly flat, which indicates that there are mostly null p-values in this region. The dotted line is at the height of the estimate of the proportion of rules that do not generate abnormal performance (i.e., $\widehat{\pi}_0$).

positive (π_A^+) and negative abnormal performance (π_A^-), which can be written as

$$\pi_A^+ = \frac{T^+(\gamma)+A^+(\gamma)}{l}, \quad \pi_A^- = \frac{T^-(\gamma)+A^-(\gamma)}{l}. \tag{4}$$

$T^+(\gamma)$ denotes the number of alternative models with positive performance and a p-value smaller than γ. $A^+(\gamma)$ denotes the number of alternative models with positive performance that are not rejected by the hypothesis test (i.e., with a p-value greater than γ). $T^-(\gamma)$ and $A^-(\gamma)$ are defined accordingly for negative performance.

Using the same approach as in Appendix B, we estimate $T^+(\gamma)$ and $T^-(\gamma)$ with

$$\widehat{T}^+(\gamma) = \widehat{R}^+(\gamma) - \widehat{F}^+(\gamma) = \{p_k \le \gamma, \varphi_k > 0;\ k = 1,\ldots,l\} - \tfrac{1}{2}\widehat{\pi}_0 l\gamma \tag{5}$$

and

$$\widehat{T}^-(\gamma) = \widehat{R}^-(\gamma) - \widehat{F}^-(\gamma) = \{p_k \le \gamma, \varphi_k < 0;\ k = 1,\ldots,l\} - \tfrac{1}{2}\widehat{\pi}_0 l\gamma. \tag{6}$$

As we increase γ, $A^+(\gamma)$ and $A^-(\gamma)$ tend to zero, while $T^+(\gamma)$ and $T^-(\gamma)$ increase. Hence, by taking a sufficiently high value γ^*, we can estimate π_A^+ and π_A^- with

$$\widehat{\pi}_A^+ = \frac{\widehat{T}^+(\gamma^*)}{l}, \quad \widehat{\pi}_A^- = \frac{\widehat{T}^-(\gamma^*)}{l}, \tag{7}$$

as explained in Barras, Scaillet, and Wermers (2010). We set $\gamma^* = 0.4$, which corresponds to the value for which $\widehat{\pi}_A^+$ and $\widehat{\pi}_A^-$ become constant.

Appendix E. Controlling the portfolio FDR$^+$ level

Storey, Taylor, and Siegmund (2004) show that the FDR point estimates can be used to define valid FDR controlling procedures under weak dependence. Hence, we can derive the following algorithm that allows the construction of a portfolio of trading rules with a FDR$^+$ level fixed at predetermined target rate. The algorithm starts with the rule having the smallest p-value (and a positive performance). Then, the rule corresponding to the next p-value is added and the FDR$^+$ recomputed. This process is repeated until we reach the desired FDR$^+$ target.

Appendix F. Determining the standard deviation of the estimators under dependence

Barras, Scaillet, and Wermers (2010) have derived the asymptotic properties of the estimators for π_0, π_A^+ and π_A^- under independent p-values. They use the large sample theory proposed by Genovese and Wasserman (2004). Here we use the results of Farcomeni (2007), who extends the results of Genovese and Wasserman (2004) to the dependent case; see also Wu (2008). The idea is to directly exploit the convergence of the empirical process associated to the p-values. In this appendix, we assume that the test statistics are totally ordered. Let us introduce the cdf $G(\lambda) = P[p_k \le \lambda]$ and its empirical counterpart $\widehat{G}(\lambda) = \sharp\{p_k \le \lambda; k = 1,\ldots,l\}/l$, $\lambda \in (0,1)$. Farcomeni (2007) shows that, when $l \to +\infty$, the empirical process $\sqrt{l}(\widehat{G}(\lambda) - G(\lambda))$ converges to a centered Gaussian random process whose covariance kernel is $K(\lambda_1,\lambda_2) = G(\min(\lambda_1,\lambda_2)) - G(\lambda_1)G(\lambda_2) + 2\sum_{j=2}^{\infty}(G_j(\lambda_1,\lambda_2) - G(\lambda_1)G(\lambda_2))$, where $G_j(\lambda_1,\lambda_2) = P[p_1 \le \lambda_1, p_j \le \lambda_2]$. That result holds true under a wide range of dependence structures such as association, latent factor model, block dependence, and mixing. Mixing refers here to spatial mixing and not temporal mixing used in the time series literature. This requires viewing the p-values corresponding to the test statistics as a spatial process on [0,1] such that the mixing conditions make the p-values located in the subintervals close to zero sufficiently independent from the p-values located in the subintervals close to one. The infinite sum in the covariance kernel corresponds to the contribution coming from dependence. We can estimate it by $2\sum_{j=2}^{a_l}(\widehat{G}_j(\lambda_1,\lambda_2) - \widehat{G}(\lambda_1)\widehat{G}(\lambda_2))$, where $\widehat{G}_j(\lambda_1,\lambda_2) = \sharp\{p_i \le \lambda_1,\ p_{i+j} \le \lambda_2;\ i = 1,\ldots,l-j\}/(l-j)$, with $a_l \to +\infty$ such that $a_l/l \to 0$. Hence, we deduce that an estimate of the standard deviation of $\widehat{\pi}_0(\lambda) = (1-\widehat{G}(\lambda))/(1-\lambda)$ under dependence is $\widehat{\sigma}_{\widehat{\pi}_0(\lambda)} = \{\widehat{G}(\lambda)(1-\widehat{G}(\lambda)) + 2\sum_{k=2}^{a_l}(\widehat{G}_k(\lambda,\lambda) - \widehat{G}(\lambda)^2)\}^{1/2}/((1-\lambda)\sqrt{l})$.

Now look at the estimator of π_A^+ (the treatment for π_A^- is similar). We recognize that, in $\widehat{\pi}_A^+ = \widehat{G}^+(\gamma^*) - (\gamma^*/2)\widehat{\pi}_0(\lambda)$, the first term $\widehat{G}^+(\gamma^*) = \widehat{R}^+(\gamma^*)/l = \sharp\{p_k \le \gamma^*, \varphi_k > 0; k = 1,\ldots,l\}/l$ is an estimate of the probability of the event $\{p_k \le \gamma^*\} \cap \{\varphi_k > 0\}$. Hence, we can estimate its standard deviation with $\widehat{\sigma}_{\widehat{G}^+(\gamma^*)} = \{\widehat{G}^+(\gamma^*)(1-\widehat{G}^+(\gamma^*)) + 2\sum_{k=2}^{a_l}(\widehat{G}_k^+(\gamma^*,\gamma^*) - \widehat{G}^+(\gamma^*)^2)\}^{1/2}/\sqrt{l}$, where $\widehat{G}_j^+(\lambda_1,\lambda_2) = \sharp\{p_i \le \lambda_1, p_{i+j} \le \lambda_2, \varphi_i > 0, \varphi_{i+j} > 0;\ i = 1,\ldots,l-j\}/(l-j)$. Combining the two results, we deduce that an estimate of the standard deviation of $\widehat{\pi}_A^+$ is $\widehat{\sigma}_{\widehat{\pi}_A^+} = \{\widehat{\sigma}^2_{\widehat{G}^+}(\gamma^*) + (\gamma^*/2)^2\widehat{\sigma}^2_{\widehat{\pi}_0(\lambda)} + 2((\gamma^*/2)/(1-\lambda))\widehat{\sigma}_{\widehat{G}^+(\gamma^*),\widehat{G}(\lambda)}\}^{1/2}/\sqrt{l}$, with the covariance term estimated by $\widehat{\sigma}_{\widehat{G}^+(\gamma^*),\widehat{G}(\lambda)} = \widehat{G}^+(\gamma^*)(1-\widehat{G}(\lambda)) + 2\sum_{k=2}^{a_l}(\widehat{G}_k^+(\gamma^*,\gamma^*) - \widehat{G}^+(\gamma^*)\widehat{G}(\lambda))$.

Appendix G. Monte Carlo experiments

We perform a simulation study showing that the FDR method correctly estimates the proportions of outperforming, underperforming, and nonperforming trading rules and that it is more powerful than the RW method. The simulations also illustrate that one explanation of the lack of power of the RW method is that it does not select further rules once it encounters a lucky rule. As the RW method is an extension of the BRC, our results simultaneously show the advantage compared with the BRC. We design the Monte Carlo simulations to match the historical performance of strategies and their empirical properties. In particular, we preserve both the time-series and the cross-sectional dependences among trading rules. By maintaining the clusters of similar strategies, e.g., filter rules or moving averages with only slightly different parameters, our Monte Carlo study illustrates the good behavior of the FDR approach even under the weak dependence structure relevant to our empirical study.

We simulate 126-day trajectories, corresponding to a six-month period, for $l = 7{,}846$ strategies as in the empirical study. We set 20% of the simulated strategies to outperform the benchmark, 50% to generate no significant abnormal returns, and 30% to deliver negative performance.

The original sample used to generate the simulated paths is a 126-day interval randomly chosen during subperiod 3 (1939–1962), to have a basis of strategies with positive performance. To generate the simulated path, we apply the stationary block bootstrap just as when computing the p-values. We do not, however, resample blocks of returns independently for each strategy. Instead, we draw $l \times b$ matrices, where b is the random size of the block in the time series dimension. This approach allows us to maintain the cross-sectional relations among same-category strategies. Because of the intrinsic properties of the stationary block bootstrap, the new paths we obtain match the empirical properties of the trading strategies, e.g., serial correlation, cross-sectional dependence, skewness, and time-varying volatility. Our simulation approach is nonparametric because we use a nonparametric bootstrap to generate the new trajectories.

To control which rules are respectively underperforming, null, and outperforming, we start by computing the average return for each simulated strategy and recenter the whole trajectory. By construction, all paths have zero mean at this step. We then select the underperforming, null, and outperforming strategies, and we shift the trajectories of the underperforming and outperforming rules, respectively, by some negative and positive value. This type of vertical parallel translation does not affect the other empirical characteristics of the trajectories because only the mean is adjusted; see Paparoditis and Politis (2003). We choose 30% of underperforming and 20% of outperforming trading rules within each of the five categories. For each category, the outperforming rules are selected as the block of adjacent rules with the highest average performance ranking in the historical sample. Underperforming rules are chosen similarly. The aim of this approach is to preserve adjacent pools of outperforming and underperforming rules. It avoids a situation in which strategies with only slightly different parameters are suddenly either outperforming or underperforming. The cross-sectional dependence among strategies is maintained.

We set the values used to shift the trajectories of the selected outperforming and underperforming rules such as to match sensible levels of Sharpe ratios corresponding to our empirical study. We choose three specific levels of outperformance, namely a positive Sharpe ratio equal to 2, 3, or 4, and three specific levels of underperformance, namely, a negative Sharpe ratio equal to -2, -3, or -4. These values correspond to annualized Sharpe ratios computed using daily returns; i.e., they are obtained by multiplying the daily mean excess return over daily standard deviation ratio by $\sqrt{252}$. Hence, the annualized Sharpe ratios of 2, 3, and 4 correspond to daily Sharpe ratio values of only 0.13, 0.19, and 0.25. In our historical sample, we observe daily Sharpe ratios as high as 0.23 ($3.6/\sqrt{252}$) for the outperforming rules and as low as -0.30 ($-4.8/\sqrt{252}$) for the underperforming rules. The positive daily Sharpe ratio of 0.13 corresponds to the 83th percentile of the distribution of observed positive daily Sharpe ratios in our sample. The negative daily Sharpe ratio of -0.13 corresponds to the 64th percentile of the distribution of observed negative daily Sharpe ratios. The annualized Sharpe ratio levels we use remain conservative, and, in particular, the $(2,-2)$ pair of outperformance versus underperformance results in a challenging setting for any rule selection method. We investigate the nine resulting combinations of specific alternative hypotheses of positive and negative Sharpe ratios, the null hypothesis being a Sharpe ratio equal to zero. This broad set of alternative hypotheses allows observing the behavior of the RW and FDR methods for outperforming and underperforming rules more or less distinguishable from rules with no genuine performance. We shift the trajectories of the different rules in such a way as to precisely obtain the same chosen positive Sharpe ratio level for all outperforming rules and the same chosen negative Sharpe ratio level for all underperforming rules. If we take as an example the pair $(2, -2)$ of Sharpe ratios for outperformance and underperformance, we construct 20% of strategies sharing the same Sharpe ratio of 2 (same

Table A1
Distribution of annualized mean excess returns corresponding to chosen Sharpe ratio levels.

This table displays the quartiles of the distribution of annualized mean excess returns (in percent) induced by setting positive and negative Sharpe ratio levels by pairs in the Monte Carlo simulations for the outperforming and underperforming strategies. The values correspond to averages over 1,000 Monte Carlo simulations. Numbers in parentheses are standard deviations. The different settings correspond to all the combinations of outperforming rules having a positive annualized Sharpe ratio (SR) equal to 2, 3, or 4, and underperforming rules having a negative annualized Sharpe ratio equal to -2, -3, or -4. The proportions of outperforming (π_A^+), underperforming (π_A^-), and zero performance (π_0) rules in the population are set to, respectively, 20%, 30%, and 50%.

Outperforming SR	Quartile	Underperforming SR					
		-2		-3		-4	
		Outperforming	Underperforming	Outperforming	Underperforming	Outperforming	Underperforming
2	1st	3.8 (2.1)	−3.2 (2.0)	3.7 (2.1)	−6.8 (2.1)	3.8 (2.1)	−10.5 (2.3)
	2nd	8.0 (3.2)	−7.5 (2.0)	7.9 (3.2)	−11.9 (2.2)	8.0 (3.2)	−16.2 (2.4)
	3rd	12.7 (4.1)	−13.3 (3.0)	12.5 (4.1)	−18.1 (2.9)	12.7 (4.2)	−23.0 (3.1)
3	1st	7.3 (2.1)	−3.3 (2.0)	7.4 (2.1)	−6.8 (2.0)	7.4 (2.1)	−10.4 (2.2)
	2nd	12.3 (3.4)	−7.6 (2.0)	12.5 (3.3)	−11.8 (2.2)	12.5 (3.3)	−16.2 (2.3)
	3rd	17.4(4.1)	−13.3 (2.8)	17.7 (4.0)	−18.0 (2.8)	17.7 (4.0)	−23.0 (2.9)
4	1st	11.0 (2.3)	−3.3 (2.0)	10.9 (2.2)	−6.9 (2.1)	10.9 (2.3)	−10.4 (2.3)
	2nd	16.7 (3.6)	−7.6 (2.0)	16.6 (3.5)	−11.9 (2.1)	16.7 (3.6)	−16.2 (2.3)
	3rd	22.5 (4.4)	−13.4 (3.0)	22.3 (4.0)	−18.2 (2.8)	22.5 (4.3)	−23.1 (2.9)

alternative hypothesis of positive performance) and 30% of strategies sharing the same Sharpe ratio of -2 (same alternative hypothesis of negative performance), the remaining 50% having a zero Sharpe ratio (same null hypothesis of zero performance). To provide an accurate idea of the alternative hypotheses, Table A1 reports the quartiles of the annualized mean excess return of the outperforming and underperforming rules, for the nine combinations of Sharpe ratios. For example, when we set the Sharpe ratio of the outperforming and underperforming rules to, respectively, 2 and -3, the annualized mean excess return lies between 3.7% and 12.5% for the outperforming rules and between -6.8% and -18.1% for the underperforming rules. Furthermore, the volatility of trading rules corresponding to the null hypothesis of zero performance is of the same order of magnitude as for the outperforming and underperforming rules.

Our results are based on 1,000 Monte Carlo iterations. Table A2 displays the estimates using the FDR method of the proportions of outperforming (π_A^+), underperforming (π_A^-), and nonperforming (π_0) rules, for the nine Sharpe ratio combinations. The reported results show that the estimates are very accurate. Unreported results obtained in a simpler setting in which strategies are all independent show that standard deviations are only slightly increased when we preserve the cross-sectional dependencies.

Next, we form portfolios of trading rules by controlling the FDR$^+$ at 10% and 20% and by using the RW approach to control the FWER at the 5% and 20% level. For the different portfolios and the nine Sharpe ratio combinations, Table A3 reports average values over the one thousand simulations for the true false discovery rate, the percentage of true outperforming rules detected, and the portfolio size. Focusing on the (3,−3) pair of Sharpe ratios (center of the table), the 10%-FDR$^+$ portfolio detects on average 52.6% of the outperforming rules and closely meets its FDR target at 9.7%. In comparison, the 5%-RW portfolio detects only 0.6% of the outperforming rules on average. Controlling the FWER at 20% with the RW approach increases the power to only 4.7%. The 20%-FDR$^+$ portfolio detects on average 64.7% of the outperforming rules. The FDR is below the target level 20% at 11.4%. This shows that the procedure achieves a control of the FDR, namely the achieved FDR is below the chosen target level as predicted by asymptotic theory, while simultaneously achieving good power properties. Hence, the FDR approach has a clear advantage over the RW method (and over the BRC) in the environment of our study. Methods based on the FWER are too conservative when l is large. The Monte Carlo study is also a good

Table A2
Average estimates of the proportions of null, outperforming, and underperforming rules, under the Sharpe ratio criterion.

This table presents the average over 1,000 Monte Carlo simulations of the false discovery rate estimates of π_0, π_A^+, and π_A^-. The true values are set respectively to 50, 20 and 30 (in percent). Numbers in parentheses correspond to standard deviations (in percent). The results are provided for the nine combinations of outperforming rules annualized Sharpe ratio (SR) set to 2, 3, or 4, and underperforming rules annualized Sharpe ratio set to -2, -3, or -4.

Outperforming SR	Proportions	Underperforming SR		
		-2	-3	-4
2	$\pi_0=50$	70.5 (6.6)	62.0 (5.4)	58.9 (6.0)
	$\pi_A^+=20$	9.8 (8.9)	9.8 (7.3)	10.0 (7.2)
	$\pi_A^-=30$	19.7 (10.5)	28.2 (7.0)	31.1 (5.2)
3	$\pi_0=50$	65.5 (6.8)	57.4 (5.7)	53.6 (5.7)
	$\pi_A^+=20$	15.3 (6.7)	15.6 (5.4)	15.7 (5.4)
	$\pi_A^-=30$	19.2 (10.1)	27.1 (7.0)	30.7 (5.1)
4	$\pi_0=50$	62.3 (7.0)	54.9 (6.1)	51.6 (6.3)
	$\pi_A^+=20$	18.2 (5.5)	17.7 (4.3)	17.9 (4.4)
	$\pi_A^-=30$	19.6 (9.3)	27.4 (6.7)	30.5 (5.3)

Table A3
Power and size of the false discovery rate (FDR) approach and of the method of Romano and Wolf (RW, 2005) under the Sharpe ratio criterion.

This table examines the composition of the 10%- and 20%-FDR$^+$ portfolio and of the 5%- and 20%-RW portfolio. It reports average values over 1,000 Monte Carlo simulations for the true false discovery rate (in %), the percentage of true outperforming rules detected (in percent), and the portfolio size. Numbers in parentheses correspond to standard deviations (in percent). The different settings correspond to all the combinations of outperforming rules having a positive annualized Sharpe ratio (SR) equal to 2, 3, or 4, and underperforming rules having a negative annualized Sharpe ratio equal to -2, -3, or -4.

Outperforming SR	Portfolio type	Underperforming SR								
		-2			-3			-4		
		FDR$^+$	Power	Portfolio size	FDR$^+$	Power	Portfolio size	FDR$^+$	Power	Portfolio size
2	10%-FDR$^+$	17.3 (11.3)	25.6 (19.5)	495 (400)	15.9 (10.5)	27.4 (20.2)	521 (404)	16.6 (10.4)	28.0 (20.5)	535 (409)
	20%-FDR$^+$	17.2 (11.1)	32.9 (22.1)	647 (480)	16.1 (10.3)	35.4 (23.0)	683 (481)	16.9 (10.2)	36.0 (23.6)	702 (492)
	5%-RW	0.8 (5.6)	0.3 (1.4)	5 (22)	0.6 (3.9)	0.2 (0.9)	4 (16)	0.3 (2.5)	0.1 (0.4)	2 (7)
	20%-RW	1.4 (5.0)	1.9 (4.3)	31 (73)	1.3 (5.4)	2.1 (4.3)	34 (77)	0.9 (4.5)	1.2 (2.8)	20 (47)
3	10%-FDR$^+$	10.4 (8.2)	48.5 (21.1)	865 (409)	9.7 (7.5)	52.6 (21.2)	924 (398)	9.9 (7.5)	53.4 (21.5)	943 (406)
	20%-FDR$^+$	11.5 (9.3)	59.9 (20.4)	1,087 (430)	11.4 (8.4)	64.7 (20.3)	1,166 (409)	11.8 (8.7)	65.7 (20.4)	1,192 (419)
	5%-RW	0.1 (0.9)	0.8 (2.4)	12 (39)	0.2 (1.8)	0.6 (1.8)	9 (28)	0.1 (0.9)	0.4 (1.2)	6 (19)
	20%-RW	0.5 (2.3)	4.9 (8.3)	78 (133)	0.3 (1.4)	4.7 (7.8)	74 (125)	0.3 (1.6)	3.5 (6.6)	55 (104)
4	10%-FDR$^+$	8.6 (8.1)	72.8 (16.4)	1,264 (328)	8.1 (7.0)	74.6 (15.0)	1,285 (290)	8.3 (7.1)	74.3 (15.6)	1,281 (301)
	20%-FDR$^+$	10.8 (9.5)	83.5 (12.2)	1,489 (294)	10.5 (8.3)	84.9 (11.5)	1,503 (257)	11.0 (8.5)	84.9 (11.9)	1,511 (269)
	5%-RW	0.0 (0.3)	2.0 (5.3)	31 (83)	0.0 (0.1)	1.7 (4.1)	26 (64)	0.0 (0.7)	1.0 (3.4)	16 (59)
	20%-RW	0.4 (2.2)	11.7 (15.0)	186 (243)	0.3 (1.9)	11.1 (14.0)	175 (223)	0.1 (0.9)	7.5 (11.2)	119 (179)

illustration for one cause behind the low power of the RW method. The RW approach starts with the best-performing rules and is not able to detect further rules once it reaches a lucky rule, i.e., a rule with no real predictive power. As our study shows, a situation in which a rule achieves one of the highest performance by luck is not uncommon, e.g., with the $(3, -3)$ pair of Sharpe ratios, the median ranking of the first lucky rule in the simulations is 422 (mean: 521, standard deviation: 386). As there are 1,569 outperforming rules in our setting, the power of the RW method cannot exceed 25% on average.

Further Monte Carlo results under a similar design, but when performance is measured with the mean return instead of the Sharpe ratio and for specific null and alternative hypotheses set in terms of mean returns, are included in the online Appendix.

Appendix H. Transaction costs and short sale constraints

Transaction costs are commonly decomposed into two major components: explicit costs and implicit costs. Explicit costs are the direct costs of trading, such as broker commissions and taxes. Implicit costs, which are harder to measure, represent indirect costs such as the price impact of the trade and the opportunity cost of failing to execute the order in a timely manner. For the period January 1991 to March 1993, Keim and Madhavan (1997) estimate that, for exchange-listed stocks, the average total cost for a buy order is 0.49% (0.31% implicit costs $+0.18\%$ explicit costs). Transaction costs were significantly more important in earlier years, particularly before commissions were deregulated in May 1975. Stoll and Whaley (1983) use published commission schedules to estimate transaction costs during the 1960–1975 period. For the largest decile of NYSE securities, they report an estimated one-way transaction cost of 1.35% (the commission plus half the bid–ask spread).

Selling short also incurs a cost. The investor willing to take a short position must borrow the stock from a current owner at a fee. In addition, other costs are associated with shorting, such as legal and institutional constraints, or the risk that the short position will have to be involuntarily closed due to recall of the stock loan (short squeeze). D'Avolio (2002), Duffie, Grleanu, and Pedersen (2002), Geczy, Musto, and Reed (2002), and Jones and Lamont (2002) provide useful analyzes of the equity loan market. While short sale costs might be low on average, they are systematically high exactly when they are critical. As for transaction costs, lending fees have declined over time. The average shorting cost in Jones and Lamont (2002) sample (1926–1933) is 35 basis points per month. For the period 2000–2001, D'Avolio (2002) reports only 41 basis points per year. However, 9% are loan market specials, with fees averaging 4.3% per annum, but reaching spectacular heights in some rare instances.

The one-way transaction costs considered in our study correspond to brokerage fees, bid–ask spread, and slippage. In practice, further costs are incurred by the manager and passed on to the investor. To have a realistic view of the current status of transaction costs and market frictions, we have contacted several retail online brokers, banks and hedge funds to gather up-to-date data. These data show that the typical levels of transaction costs offsetting out-of-sample performance found in our study can be viewed as low (conservative). In recent periods, most of the time we do not detect genuine performance already before transaction costs (see Table 5).

We consider three real-life cases for a fund manager trading on futures. Trading an exchange traded fund instead of futures further raises the trading costs. Besides, index futures are sufficiently liquid nowadays to avoid any price impact when the traded volume stays below several millions of dollars (based on the current market liquidity conditions). This does not necessarily apply in the case of ETFs. For futures, it is considered that trading up to 2% of the average daily volume (computed over 20 days typically) does not lead to a price impact. The average daily number of E-mini S&P 500 futures traded on the Chicago Mercantile Exchange is around 2.5 million contracts. This corresponds to a daily traded volume of USD 148 billions, and the price should not be impacted if we trade below USD 3 billion. For the E-mini Dow, we have 80,000 contracts corresponding to USD 4.5 billion. This gives a limit of USD 100 million.

Table A4
Total expense ratios for managed account, off-shore fund, and on-shore fund.

This table decomposes the components of the total expense ratio for three types of vehicles: a managed account, an off-shore fund, and an on-shore fund. The four components correspond to the cost of the structure, the custody and administration costs, the transaction costs, and the management fees. The sum of the three first component makes the total operating costs. We consider four levels of asset under management: USD 0.1, 1, 10, and 100 million. All costs are listed as basis points per year.

	Asset under management (USD million)											
	Managed account				Off-shore fund				On-shore fund			
	0.1	1	10	100	0.1	1	10	100	0.1	1	10	100
Cost of structure	–	–	–	–	5	5	5	5	60	60	60	60
Custody & admin.	30	21	10	5	30	21	10	5	40	31	20	5
Transaction costs	26	26	26	26	26	26	26	26	26	26	26	26
Total operating costs	56	47	36	31	61	52	41	36	126	117	106	91
Management fees	200	200	200	200	200	200	200	200	200	200	200	200
Total expense ratio (TER)	256	247	236	231	261	252	241	236	326	317	306	291

We investigate all types of costs incurred through a managed account, an off-shore fund, or a on-shore fund. The managed account case is not far from a retail investor trading by himself through an online broker. The off-shore fund case corresponds to a fund manager living within a light and unregulated structure. The on-shore fund case concerns a fund regulated by a supervisory authority. The types of costs cover the cost of the structure (the vehicle), the custody and administration costs, and the brokerage fees. Common costs driven by the specific selected structure include formation expenses (amortization of fund creation expenses), legal expenses (audit, fund prospectus), taxes, and hiring of a management company. Custody and administration costs relate to expenses induced by the deposit of the assets and their administration (net asset value computation, reporting, hedging) as well as the booking fees (fees charged by the custodian for each executed transaction). We report transaction costs corresponding to a low turnover, namely, 20 times the asset under management (AUM) per year. In the case of our study, this amounts to buying or selling 20 times the index based on 20 trading signals. These costs include explicit and implicit costs, namely, brokerage fees, bid–ask spread, and slippage. Beside these costs linked to fund operations, we need to add management fees (fund manager remuneration) here taken as a standard 2% flat without a performance fee. The sum of all these expenses and fees make up the so-called total expense ratio (TER). Table A4 shows that the costs before accounting for management fees, i.e., the total operating costs, currently range from 31 basis points to 126 basis points per year, and that the final TER ranges from 231 basis points to 326 basis points per year. These figures are computed from a span of four values for the AUM: USD 0.1, 1, 10, and 100 millions.

Appendix I. Supplementary data

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jfineco.2012.06.001.

References

Abramovich, F., Benjamini, Y., Donoho, D., Johnstone, I., 2006. Adapting to unknown sparsity by controlling the false discovery rate. Annals of Statistics 34, 584–653.

Allen, F., Karjalainen, R., 1999. Using genetic algorithms to find technical trading rules. Journal of Financial Economics 51, 245–271.

Barras, L., Scaillet, O., Wermers, R., 2010. False discoveries in mutual fund performance: measuring luck in estimated alphas. Journal of Finance 65, 179–216.

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

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Annals of Statistics 29, 1165–1188.

Bessembinder, H., Chan, K., 1998. Market efficiency and the returns to technical analysis. Financial Management 27, 5–17.

Blanchet-Scalliet, C., Diop, A., Gibson, R., Talay, D., Tanr, E., 2007. Technical analysis compared to mathematical models based methods under parameters mis-specification. Journal of Banking and Finance 31, 1351–1373.

Brennan, M., Schwartz, E., Lagnado, R., 1997. Strategic asset allocation. Journal of Economic Dynamics and Control 21, 1377–1403.

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

Brown, S., Goetzmann, W., 1995. Performance persistence. Journal of Finance 50, 679–698.

Brown, S., Goetzmann, W., Kumar, A., 1998. The Dow theory: William Peter Hamilton's track record reconsidered. Journal of Finance 53, 1311–1333.

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

D'Avolio, G., 2002. The market for borrowing stock. Journal of Financial Economics 66, 271–306.

Diebold, F.X., 2006. Elements of Forecasting. South-Western College Pub, Cincinnati.

Duffie, D., Grleanu, N., Pedersen, L.H., 2002. Securities lending, shorting, and pricing. Journal of Financial Economics 66, 307–339.

Elliott, G., Timmermann, A., 2008. Economic forecasting. Journal of Economic Literature 46, 3–56.

Fama, E., Blume, M., 1966. Filter rules and stock-market trading. Journal of Business 39, 226–241.

Farcomeni, A., 2007. Some results on the control of the false discovery rate under dependence. Scandinavian Journal of Statistics 34, 275–297.

Finner, H., Roters, M., 2002. Multiple hypotheses testing and expected number of type I errors. Annals of Statistics 30, 220–238.

Friedman, B., 1996. Economic implications of changing share ownership. Journal of Portfolio Management 22, 59–70.

Fung, W., Hsieh, D.A., 1997. Empirical characteristics of dynamic trading strategies: the case of hedge funds. Review of Financial Studies 10, 275–302.

Gatev, E., Goetzmann, W., Rouwenhorst, G., 2006. Pairs trading: performance of a relative-value arbitrage rule. Review of Financial Studies 19, 797–827.

Geczy, C., Musto, D., Reed, A., 2002. Stocks are special too: an analysis of the equity lending market. Journal of Financial Economics 66, 241–269.

Genovese, C., Wasserman, L., 2004. A stochastic process approach to false discovery control. Annals of Statistics 32, 1035–1061.

Gompers, P., Metrick, A., 2001. Institutional investors and equity prices. Quarterly Journal of Economics 116, 229–259.

Hansen, P., 2005. A test for superior predictive ability. Journal of Business and Economic Statistics 23, 365–380.

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

Hsu, P.-H., Kuan, C.-M., 2005. Reexamining the profitability of technical analysis with data snooping checks. Journal of Financial Econometrics 3, 606–628.

Huang, W., Liu, Q., Rhee, S.G., Zhang, L., 2010. Return reversals, idiosyncratic risk, and expected returns. Review of Financial Studies 23, 147–168.

Jacquier, E., Yao, T., 2002. Re-evaluating dynamic trading strategies: the free lunch was no banquet. Unpublished working paper. University of Montreal, Montreal, Canada.

Jegadeesh, N., 1990. Evidence of predictable behavior of security returns. Journal of Finance 45, 881–898.

Jondeau, E., Rockinger, M., 2006. Optimal portfolio allocation under higher moments. European Financial Management 12, 29–55.

Jones, C., Lamont, O., 2002. Short-sale constraints and stock returns. Journal of Financial Economics 66, 207–239.

Kavajecz, K., Odders-White, E., 2004. Technical analysis and liquidity provision. Review of Financial Studies 17, 1043–1071.

Keim, D., Madhavan, A., 1997. Transactions costs and investment style: an inter-exchange analysis of institutional equity trades. Journal of Financial Economics 46, 265–292.

Kosowski, R., Naik, N., Teo, M., 2007. Do hedge funds deliver alpha? A Bayesian and bootstrap analysis. Journal of Financial Economics 84, 229–264.

Lo, A., MacKinlay, A., 1990. Data snooping biases in tests of financial asset pricing models. Review of Financial Studies 3, 431–467.

Lo, A., Mamaysky, H., Wang, J., 2000. Foundations of technical analysis: computational algorithms, statistical inference, and empirical implementation. Journal of Finance 55, 1705–1765.

Menkhoff, L., Taylor, M., 2007. The obstinate passion of foreign exchange professionals: technical analysis. Journal of Economic Literature 45, 936–972.

Neely, C., Weller, P., Dittmar, R., 1997. Is technical analysis in the foreign exchange market profitable? A genetic programming approach. Journal of Financial and Quantitative Analysis 32, 405–426.

Neftci, S., 1991. Naive trading rules in financial markets and Wiener–Kolmogorov prediction theory: a study of technical analysis. Journal of Business 64, 549–571.
Paparoditis, E., Politis, D., 2003. Residual-based block bootstrap for unit root testing. Econometrica 71, 813–855.
Politis, D., Romano, J., 1994. The stationary bootstrap. Journal of the American Statistical Association 89, 1303–1313.
Ready, M., 2002. Profits from technical trading rules. Financial Management 31, 43–61.
Romano, J., Shaikh, A., Wolf, M., 2008a. Control of the false discovery rate under dependence using the bootstrap and subsampling. Test 17, 417–442.
Romano, J., Shaikh, A., Wolf, M., 2008b. Formalized data snooping based on generalized error rates. Econometric Theory 24, 404–447.
Romano, J., Wolf, M., 2005. Stepwise multiple testing as formalized data snooping. Econometrica 73, 1237–1282.
Stoll, H., Whaley, R., 1983. Transaction costs and the small firm effect. Journal of Financial Economics 12, 57–79.
Storey, J., 2002. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B 64, 479–498.
Storey, J., 2003. The positive false discovery rate: a Bayesian interpretation and the q-value. Annals of Statistics 31, 2013–2035.
Storey, J., Taylor, J., Siegmund, D., 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. Journal of the Royal Statistical Society, Series B 66, 187–205.
Storey, J., Tibshirani, R., 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences 100, 9440–9445.
Sullivan, R., Timmermann, A., White, H., 1999. Data snooping, technical trading rule performance, and the bootstrap. Journal of Finance 54, 1647–1691.
White, H., 2000. A reality check for data snooping. Econometrica 68, 1097–1126.
Wu, W., 2008. On false discovery control under dependence. Annal of Statistics 36, 364–380.