Reproducibility and Open Science in Empirical Finance

Dennis Hoffmann

Abstract

The reproducibility of empirical results in finance has come under sustained scrutiny. A 2016 survey of 1,500 scientists across disciplines found that more than 70% had tried and failed to reproduce another researcher's experiments (Baker, 2016), and finance has not been exempt from these concerns. The "factor zoo" problem, in which hundreds of published asset-pricing factors compete for empirical support, has raised fundamental questions about the reliability of the empirical finance literature (Harvey et al., 2016). This article surveys the evidence on replication failures and their methodological causes (p-hacking, data snooping, publication bias), examines the open science practices emerging in response, and considers the unique tension between transparency ideals and finance's dependence on proprietary data. It concludes by discussing implications for machine learning and digital finance research.

1 Introduction

The question of whether published empirical results can be independently reproduced has become one of the defining methodological concerns of modern science. Baker (2016) documented widespread anxiety about reproducibility across the natural and social sciences, with a majority of surveyed researchers reporting that the scientific community faces a "significant" or "slight" reproducibility crisis. In finance, these concerns have a specific and consequential manifestation: the proliferation of published asset-pricing factors, trading strategies, and return predictors, many of which appear to lose their predictive power once subjected to rigorous out-of-sample testing or independent replication.

Harvey (2017), in his American Finance Association Presidential Address, argued that the field of financial economics needs to raise its scientific standards. He proposed that researchers adopt higher statistical thresholds, report all tested specifications, and treat the multiple-testing problem as a first-order concern rather than an afterthought. This call reflected a growing recognition that standard practices in empirical finance, including selective reporting, flexible model specification, and inadequate correction for data mining, may have inflated the proportion of false discoveries in the published literature.

This article surveys the current state of reproducibility and open science in empirical finance. Section 2 reviews the replication crisis, focusing on the factor zoo and the results of large-scale replication studies. Section 3 examines the methodological mechanisms that produce false positives: data snooping, p-hacking, and publication bias. Section 4 discusses open science practices and the tension between transparency and proprietary data. Section 5 considers implications for machine learning and digital finance. Section 6 concludes.

2 The Replication Crisis in Finance

2.1 The Factor Zoo

The term "factor zoo" describes the rapid accumulation of published asset-pricing factors. Harvey et al. (2016) catalogued 316 factors that had been proposed in the academic literature as predictors of cross-sectional stock returns. They argued that the conventional significance threshold of $t > 2.0$ is far too permissive when hundreds of factors have been tested, because the probability of at least one spurious discovery rises sharply with the number of tests. Their analysis suggested that a threshold closer to $t > 3.0$ would be more appropriate after accounting for the cumulative multiple-testing burden.

The challenge of distinguishing genuine risk factors from statistical artifacts has motivated several methodological advances. Feng et al. (2020) developed a double-selection procedure that tests whether a new factor provides incremental explanatory power beyond a set of established factors, addressing the problem that many proposed factors are redundant or subsumed by existing ones. Bryzgalova et al. (2023) took a Bayesian approach, evaluating the posterior probability of factor inclusion across a vast model space (on the order of two quadrillion models), and found that only a small number of factors survive rigorous model comparison. Barillas and Shanken (2018) developed a Bayesian framework for comparing asset-pricing models, showing that model comparison depends only on the factors that differ between competing specifications.

2.2 Notable Replication Studies

Several large-scale replication studies have directly assessed the robustness of published anomalies and return predictors. Hou et al. (2020) attempted to replicate 452 anomalies documented in the literature and found that the majority, roughly 65%, could not be replicated at conventional significance levels. Their analysis identified both data-processing choices and sample-period sensitivity as important sources of non-replication. McLean and Pontiff (2016) examined whether the returns to published anomalies decay after academic publication and found statistically significant post-publication declines, consistent with the hypothesis that some portion of published effects reflects data mining rather than genuine economic phenomena.

Jensen et al. (2023) conducted the most geographically comprehensive replication study to date, testing 153 factors across 93 countries. Their findings were nuanced: many factors do replicate in the sense that they retain the predicted sign, but their economic magnitudes are often substantially reduced relative to the original studies. This pattern suggests that while the factor zoo contains real phenomena, the published effect sizes are systematically inflated. Chordia et al. (2020) demonstrated that a substantial fraction of documented anomalies disappear entirely when proper multiple-testing corrections are applied, reinforcing concerns about the reliability of uncorrected significance tests. Goyal and Jegadeesh (2018) provided additional evidence by examining the distinction between cross-sectional and time-series tests of return predictability, finding that results can be highly sensitive to the choice of testing framework.

3 P-Hacking, Data Snooping, and Publication Bias

3.1 Data Snooping and Multiple Testing

Data snooping occurs when a researcher conducts many tests on the same dataset and selectively reports the most favourable results. In empirical finance, where researchers routinely search across factors, time periods, subsamples, and model specifications, the scope for data snooping is substantial. Sullivan et al. (1999) provided an early and influential treatment of this problem in the context of technical trading rules. They developed bootstrap-based tests that account for the full universe of trading rules evaluated (not just the reported "best" rule), demonstrating that many seemingly profitable strategies lose their significance once data snooping is properly controlled.

P-hacking is a closely related practice in which researchers exploit flexibility in data analysis, such as choosing among multiple dependent variables, control variables, sample definitions, or functional forms, to obtain statistically significant results. HARKing (Hypothesizing After the Results are Known) compounds this problem by allowing researchers to present post hoc findings as though they were predicted ex ante, obscuring the exploratory nature of the analysis. Harvey (2017) emphasized that these practices are particularly damaging in finance because the economic stakes of false discoveries are high: investment strategies built on spurious factors can result in significant capital misallocation.

The reproducibility crisis is not unique to finance. Hensel (2021) compared the diagnosis and response to reproducibility concerns across management, psychology, and economics, finding that while each field has identified similar problems, the pace and nature of institutional responses differ considerably. Finance has been slower than psychology to adopt formal remedies such as preregistration, though awareness of the issue has grown rapidly since the mid-2010s.

3.2 Publication Bias and Lucky Factors

Publication bias, the tendency for journals to publish statistically significant results over null findings, systematically distorts the published evidence base. Harvey and Liu (2021) formalized this concern through their analysis of "lucky factors": factors that appear significant purely by chance when many candidates are tested simultaneously. They developed a framework for estimating the proportion of published factors that are likely to be false discoveries, finding that the rate of spurious factors is non-trivial under plausible assumptions about the number of tests conducted across the research community.

Chen and Zimmermann (2020) quantified publication bias in the cross-section of stock returns by comparing the distribution of reported t-statistics to the distribution expected under the null hypothesis. Their analysis revealed clear evidence of selective reporting around conventional significance thresholds, with a marked excess of results just above $t = 1.96$. Mitton (2022) extended the analysis of methodological flexibility to corporate finance, documenting how researchers' choices among plausible specifications can shift results from insignificant to significant, and vice versa.

Maniadis et al. (2017) provided a theoretical framework for understanding how replication power, prior beliefs, and publication incentives interact to determine the credibility of

research findings. Their model predicts that when priors are diffuse and replication is rare, the false discovery rate among published results can be alarmingly high, even if individual studies are conducted in good faith. This analysis underscores the importance of systematic replication as a corrective mechanism.

4 Open Science Practices and the Proprietary Data Tension

4.1 Open Science Practices in Finance

Open science encompasses a set of practices designed to make the research process more transparent and its outputs more accessible. The core practices include preregistration (specifying hypotheses, data, and analysis plans before conducting a study), open data (sharing the datasets used in analysis), open code (publishing analysis scripts and software), and registered reports (submitting study designs for peer review before results are known).

The landmark effort by the Open Science Collaboration (2015) to estimate the reproducibility of psychological science, in which 100 published studies were independently replicated, demonstrated both the feasibility and the value of large-scale replication projects. Only about 36% of the original findings replicated at conventional significance levels, a result that galvanized the broader scientific community and accelerated the adoption of open science norms across disciplines. In finance, adoption of these practices has been comparatively slow. Preregistration remains rare in top finance journals, and registered reports have seen near-zero uptake, partly because the culture of finance research emphasizes novelty and theoretical contribution over procedural transparency (Baker, 2016).

4.2 The Proprietary Data Problem

A distinctive challenge for open science in finance is the field's dependence on proprietary and licensed databases. Most empirical work in asset pricing relies on data from the Center for Research in Security Prices (CRSP) and Compustat, accessed through Wharton Research Data Services (WRDS). High-frequency and microstructure studies typically use the Trade and Quote (TAQ) database. Global and fixed-income research often depends on Bloomberg or Refinitiv terminals. These databases are expensive, institutionally licensed, and subject to terms of service that prohibit redistribution.

This creates a fundamental tension between open science ideals and research practice. A researcher can share code, describe methodology in detail, and preregister hypotheses, but cannot share the underlying data without violating licensing agreements. The result is an asymmetry in replication ability: researchers at well-resourced institutions with database subscriptions can attempt replications, while those without access cannot. Pérignon et al. (2024) highlighted how this institutional barrier compounds other reproducibility challenges, noting that even when code and data are nominally available, differences in database vintages and access levels can prevent exact reproduction.

Emerging solutions include the creation of synthetic or simulated datasets that preserve the statistical properties of proprietary data without revealing individual observations,

controlled-access data repositories that allow replication under restricted conditions, and journal policies requiring data-access statements that specify exactly which databases and vintages were used. None of these fully resolves the tension, but each represents a step toward reconciling proprietary constraints with reproducibility goals.

4.3 Computational Reproducibility

Computational reproducibility, the ability to reproduce a study's results given the same data and code, is a necessary (though not sufficient) condition for scientific credibility. Pérignon et al. (2024) organized a large-scale reproducibility exercise in which independent research teams attempted to computationally reproduce results from papers published in leading finance journals, generating approximately 1,000 reproduction tests. They found that exact reproduction was frequently hindered by undocumented data transformations, differences in software versions, random-number seeds, or platform-specific numerical behaviour, even when code was nominally available.

Koenker and Zeileis (2009) argued early on that reproducible econometric research requires a commitment to open-source tools, literate programming, and the archiving of complete computational environments. Their vision has been partially realized through the growing use of version-controlled repositories, containerized environments (such as Docker), and literate programming frameworks (such as Jupyter notebooks and R Markdown).

Chen and Zimmermann (2022) provided a landmark example of open infrastructure for finance by constructing an open-source repository that replicates over 200 cross-sectional return predictors. Their `openassetpricing.com` platform allows researchers to download replicated factor portfolios, compare methodological choices across studies, and build on a common, transparent foundation. This project demonstrates that large-scale open science in finance is technically feasible, even within the constraints of proprietary data, when the focus is on replicable methodology rather than raw data sharing.

5 Implications for Machine Learning and Digital Finance

The application of machine learning methods to financial prediction introduces both new reproducibility challenges and new opportunities for transparent research practice. Kapoor and Narayanan (2023) conducted a systematic review of data leakage across 17 scientific fields and found that leakage (the inadvertent use of test-set information during model training) was present in a substantial fraction of published ML studies, affecting the reported performance of 294 papers. In financial applications, data leakage can take discipline-specific forms, such as using future information in feature construction, failing to respect the temporal ordering of training and test sets, or applying standard cross-validation procedures that are inappropriate for time-series data.

Gu et al. (2020) demonstrated how rigorous machine learning methodology can be applied to asset pricing. Their study of empirical asset pricing via machine learning employed a comprehensive set of predictors, strict out-of-sample evaluation, and transparent reporting of model selection procedures. By establishing a benchmark for methodological transparency in ML-based finance research, their work illustrates that high-dimensional methods and

reproducible practice are not inherently in conflict. López de Prado (2018) provided practical guidance on building reproducible ML pipelines for financial applications, emphasizing the importance of combinatorial purged cross-validation, proper backtesting protocols, and the avoidance of common pitfalls such as overfitting to historical data.

Digital finance, including fintech, algorithmic trading, and blockchain-based systems, creates new contexts for these reproducibility concerns. Alternative data sources (satellite imagery, social media sentiment, web-scraped data) raise questions about data provenance, versioning, and long-term availability that traditional financial databases do not. At the same time, digital finance offers tools that can advance reproducibility: version-controlled analysis pipelines, containerized computational environments, and automated testing frameworks are native to software engineering practice and increasingly adopted in quantitative finance. The challenge is to bridge the gap between the software engineering culture of reproducibility by default and the academic finance culture in which reproducibility has traditionally been an afterthought.

Open-source ML frameworks, standardized benchmarking datasets, and community-driven replication efforts represent concrete steps in this direction. Chen and Zimmermann (2022) demonstrated with their open-source factor repository that transparent, community-maintained research infrastructure is viable in finance. Extending this model to ML-based predictions and digital finance applications is a natural and necessary next step.

6 Conclusion

The reproducibility of empirical results in finance has improved in important respects since the concerns raised by Harvey et al. (2016) and Harvey (2017) entered mainstream discourse. Large-scale replication studies have provided the field with a clearer picture of which findings are robust and which are fragile. Statistical methods for addressing multiple testing, data snooping, and publication bias have matured considerably. Open-source tools and platforms such as `openassetpricing.com` have demonstrated that transparent, community-maintained research infrastructure is feasible in finance.

Significant barriers remain, however. The dependence on proprietary databases creates structural obstacles to full data sharing that have no simple solution. Preregistration and registered reports remain rare in finance, and publication incentives continue to favour novelty over replication. The growing use of machine learning methods introduces new reproducibility risks (data leakage, opaque model complexity, inadequate temporal validation) even as it expands the toolkit for transparent, automated research pipelines.

Addressing these challenges will require coordinated action across multiple levels: individual researchers adopting transparent practices, journals enforcing reproducibility policies, institutions supporting data-sharing agreements, and the research community investing in open infrastructure. The evidence reviewed in this article suggests that the tools and methods for reproducible finance research exist; the remaining challenge is one of adoption and institutional change.

References

Monya Baker. 1,500 scientists lift the lid on reproducibility, 2016.

Francisco Barillas and Jay Shanken. Comparing asset pricing models. *Journal of Finance*, 73(2):715–754, 2018. doi: 10.1111/jofi.12607.

Svetlana Bryzgalova, Jiantao Huang, and Christian Julliard. Bayesian solutions for the factor zoo: We just ran two quadrillion models. *Journal of Finance*, 78(1):487–557, 2023. doi: 10.1111/jofi.13197.

Andrew Y. Chen and Tom Zimmermann. Publication bias and the cross-section of stock returns. *Review of Asset Pricing Studies*, 10(2):249–289, 2020. doi: 10.1093/rapstu/raz011.

Andrew Y. Chen and Tom Zimmermann. Open source cross-sectional asset pricing. *Critical Finance Review*, 11(2):207–264, 2022. doi: 10.1561/104.00000112.

Tarun Chordia, Amit Goyal, and Alessio Saretto. Anomalies and false rejections. *Review of Financial Studies*, 33(5):2134–2179, 2020. doi: 10.1093/rfs/hhaa018.

Guanhao Feng, Stefano Giglio, and Dacheng Xiu. Taming the factor zoo: A test of new factors. *Journal of Finance*, 75(3):1327–1370, 2020. doi: 10.1111/jofi.12883.

Amit Goyal and Narasimhan Jegadeesh. Cross-sectional and time-series tests of return predictability: What is the difference? *Review of Financial Studies*, 31(5):1784–1824, 2018. doi: 10.1093/rfs/hhx131.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5):2223–2273, 2020. doi: 10.1093/rfs/hhaa009.

Campbell R. Harvey. Presidential address: The scientific outlook in financial economics. *Journal of Finance*, 72(4):1399–1440, 2017. doi: 10.1111/jofi.12530.

Campbell R. Harvey and Yan Liu. Lucky factors. *Journal of Financial Economics*, 141(2):413–435, 2021. doi: 10.1016/j.jfineco.2021.04.014.

Campbell R. Harvey, Yan Liu, and Heqing Zhu. . . . and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68, 2016. doi: 10.1093/rfs/hhv059.

Przemyslaw G. Hensel. Reproducibility and replicability crisis: How management compares to psychology and economics – a systematic review. *European Management Journal*, 39(5):577–594, 2021. doi: 10.1016/j.emj.2021.01.002.

Kewei Hou, Chen Xue, and Lu Zhang. Replicating anomalies. *Review of Financial Studies*, 33(5):2019–2133, 2020. doi: 10.1093/rfs/hhy131.

Theis Ingerslev Jensen, Bryan T. Kelly, and Lasse Heje Pedersen. Is there a replication crisis in finance? *Journal of Finance*, 78(5):2465–2518, 2023. doi: 10.1111/jofi.13249.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804.

Roger Koenker and Achim Zeileis. On reproducible econometric research. *Journal of Applied Econometrics*, 24(5):833–847, 2009. doi: 10.1002/jae.1083.

Marcos López de Prado. *Advances in Financial Machine Learning*. Wiley, 2018. doi: 10.1002/9781119482086.

Zacharias Maniadis, Fabio Tufano, and John A. List. To replicate or not to replicate? exploring reproducibility in economics through the lens of a model and a pilot study. *Economic Journal*, 127(605):F209–F235, 2017. doi: 10.1111/ecoj.12527.

R. David McLean and Jeffrey Pontiff. Does academic research destroy stock return predictability? *Journal of Finance*, 71(1):5–32, 2016. doi: 10.1111/jofi.12365.

Todd Mitton. Methodological variation in empirical corporate finance. *Review of Financial Studies*, 35(2):527–575, 2022. doi: 10.1093/rfs/hhab030.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015. doi: 10.1126/science.aac4716.

Christophe Pérignon, Olivier Music Akmansoy, Christophe Hurlin, Anna Dreber, Zuzana Irsova, Magnus Johannesson, and Michael Kirchler. Computational reproducibility in finance: Evidence from 1,000 tests. *Review of Financial Studies*, 2024. doi: 10.1093/rfs/hhae029.

Ryan Sullivan, Allan Timmermann, and Halbert White. Data-snooping, technical trading rule performance, and the bootstrap. *Journal of Finance*, 54(5):1647–1691, 1999. doi: 10.1111/0022-1082.00163.