Empirical Economics (2026) 70:63
https://doi.org/10.1007/s00181-026-02905-6

The permutation test for event studies with a small number of events

Phuong Anh Nguyen[1,2] · **Michael Wolf**[3,4]

Received: 17 June 2025 / Accepted: 20 February 2026

Abstract
Return event studies typically involve a large number of event instances. In some applications, however, this number may be very small–sometimes as few as two event instances. In such cases, standard approaches to testing average abnormal returns (AAR) or cumulative average abnormal returns (CAAR) are less effective, or may not apply at all, as they rely on central limit theorems that require large sample sizes. We propose a nonparametric permutation test that remains valid for arbitrarily small numbers of event instances. Its performance is evaluated via Monte Carlo studies, and the method is further illustrated using two empirical applications.

Keywords Cumulative average abnormal return · Event study · Permutation test

JEL Classification C12 · G14

1 Introduction

Return event studies have many applications in accounting and finance; for example, see Campbell et al. (1997, Chapter 4), MacKinlay (1997), Kothari and Warner (2007), Kliger and Gurevich (2014), and the references therein. Given the (intended) brevity of this paper, we assume that readers have basic familiarity with the field; otherwise they should feel free to consult the listed references first.[1]

Return event studies examine whether abnormal returns on an event date—or, more generally, within an event window—are unusually large in magnitude. This question

[1] Apart from return event studies there are also trading-volume event studies and volatility event studies, but those are not the topic of this paper.

✉ Michael Wolf
michael.wolf@econ.uzh.ch

1 International University, VNU-HCM, Ho Chi Minh City, Vietnam

2 Viet Nam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam

3 Department of Economics, University of Zurich, Zürich, Switzerland

4 ADIA Lab, Abu Dhabi, United Arab Emirates

Published online: 24 March 2026

is addressed through a formal hypothesis test in which the null hypothesis specifies that the expected value of a suitably defined abnormal-return variable equals zero. Rejection of the null implies that the event had an impact on (average) returns.

Throughout this paper, we use the term event instance to denote the basic unit of observation in an event study. An event instance is defined as a specific combination of a firm and an event date (or, more generally, an event window). Thus, a single event affecting multiple firms gives rise to multiple event instances, as does a single firm affected by the same type of event at multiple dates. We denote by F the total number of event instances considered in the analysis.

In the simplest case of a single firm and a single event time ("single-firm, single-event study," $F = 1$), the relevant random variable is either the abnormal return on the event date (AR) or the cumulative abnormal return over an event window (CAR). When multiple event instances are considered ($F > 1$), these quantities are averaged across event instances, yielding the average abnormal return on event days[2] (AAR) or the average cumulative abnormal return over event windows, commonly referred to as the cumulative average abnormal return (CAAR).

When testing AAR or CAAR, the number of event instances, F, effectively serves as the sample size for deriving the (approximate) sampling distribution of the test statistic under the null hypothesis. When F is large, often a central limit theorem can be invoked and inference can be conducted using standard parametric tests.[3] When F is smaller—but not too small—often nonparametric tests can be applied. Even these tests, however, require a minimal sample size; although there is no sharp cutoff, a practical lower bound is often on the order of ten event instances.

As an illustration, consider the sign test for testing AAR and assume that each firm is affected by the event only once, so that F equals the number of firms under study. The test statistic is the number of firms with a positive AR on the event day. Let I_f denote an indicator variable that equals one if firm f has a positive AR on the event date. Under the null hypothesis that the I_f are independent and identically distributed (i.i.d.) Bernoulli random variables with success probability $p = 0.5$, the test statistic follows a $Bin(F, 0.5)$ distribution, a binomial distribution with parameters F and 0.5. For a one-sided test, the smallest attainable p-value is therefore 0.5^F, corresponding to the probability that all ARs are positive; for a two-sided test, the smallest attainable p-value is 0.5^{F-1}, the probability that all ARs have the same sign.[4] Consequently, at the conventional 5% significance level, the null hypothesis can never be rejected if $F \leq 4$ for the one-sided test, or if $F \leq 5$ for the two-sided test. Thus, the sign test becomes uninformative when the number of firms—or, more generally, the number of event instances— is sufficiently small.

[2] More generally, this can refer to the average abnormal return on any specific day within the event window, such as the day after the event took place.

[3] The term "parametric test" is somewhat of a misnomer in this context, as it does not require assuming that abnormal returns follow a specific parametric family, such as the normal distribution. Rather, the term reflects that the approximate null distribution of the test statistic is taken to be parametric, for example a standard normal or a t-distribution with certain degrees of freedom.

[4] Under the (reasonable) assumption that ARs are continuous random variables, the probability that an AR equals zero is zero.

Similar limitations apply to other nonparametric tests, including the generalized sign test, the Corrado rank test, the generalized rank t-test, and the Wilcoxon signed-rank test; for example, see (Cowan 1992; Kolari and Pynnonen 2011), and Wilcoxon (1945). As an aside, when the sign test or the generalized sign test is implemented using a normal approximation rather than the exact Binomial distribution—as is often done in practice—at least $F = 20$ firms are required for the approximation to be reliable.[5]

The goal of this paper is to propose tests that remain informative even when the number of event instances is very small, potentially as small as two. In such settings, meaningful inference requires that sufficiently strong evidence against the null hypothesis can, in principle, yield very small p-values. We achieve this by adapting the general framework of permutation tests to the problem of testing AAR and CAAR. Related permutation-based procedures for testing AR and CAR in the single-firm, single-event case have previously been considered by Nguyen and Wolf (2024). Our contribution can therefore be viewed as an extension of their approach to settings with multiple event instances.

One may reasonably question whether it makes sense to conduct an event study when the number of affected firms is very small.[6] In particular, it may seem problematic to extrapolate evidence from such a study—for example, concerning the impact of a regulatory shock or a specific corporate announcement—from only a few firms to a broader population.

We offer three observations in this regard. First, one may always analyze the impact on each firm individually. In that case, existing firm-level event-study methodologies apply; for example, the test proposed by Nguyen and Wolf (2024) can be used on a firm-by-firm basis when each firm is affected by the event only once. Second, in some markets there may exist only a small number of comparable events–such as regulatory interventions or industry-specific announcements—affecting only a few firms. In such settings, inference on the average effect across the affected firms may be more informative than a collection of individual case studies. Third, a single firm may be affected by the same type of event repeatedly, but only a small number of times. In this case as well, inference on the average effect remains of interest, and the methodology proposed in this paper is directly applicable.

Taken together, these considerations highlight the value of exact, finite-sample inference for (cumulative) average abnormal returns when the number of event instances is small. Our approach does not rely on asymptotic approximations and therefore remains valid in settings where standard large-sample methods are not applicable. Expanding the methodological toolbox available to event-study researchers is particularly valuable in precisely those settings where conventional approaches break down.

The remainder of the paper is structured as follows. Section 2 states the problem formulation. Section 3 proposes an alternative test statistic which is more suitable for our purposes than the classic test statistic for testing CAAR. Section 4 details

[5] As a textbook rule, one needs $\min\{Fp, F(1-p)\} \geq 10$ for the normal approximation to the Binomial to be trustworthy; when $p = 0.5$, this rule results in $F \geq 20$.

[6] We thank a referee for raising this point and for helpful perspective.

the proposed permutation test. Section 5 addresses finite-sample performance via two Monte Carlo studies. Section 6 provides an application of the proposed test to two real-life examples. Section 7 concludes.

2 Problem formulation

Consider a generic event instance, that is, a combination of a firm and an event date. Abnormal returns are computed using a chosen expected-return model, such as the constant-mean model, the market model, the CAPM, or a multi-factor model; our methodology is agnostic with respect to this choice. Returns are indexed in event time by t (though some authors use τ). Let $t = 0$ denote the event date, and let the event window range from $t = T_1 + 1$ to $t = T_2$, with size $m := T_2 - T_1$; naturally, $T_1 + 1 \leq 0 \leq T_2$. To unify the exposition, we allow the case $T_1 + 1 = 0 = T_2$, where the event window consists only of the event day ($m = 1$). The estimation window spans from $t = T_0 + 1$ to $t = T_0 + n \leq T_1$. A leading case in the literature is $T_0 + n = T_1$, where the estimation window ends immediately before the event window begins; see MacKinlay (1997, Section 5). We do not impose this condition here, as in practice a gap between the two windows may be desirable. With AR_t denoting the abnormal return of the firm on day t, the cumulative abnormal return during the event window is given by

$$CAR := \sum_{t=T_1+1}^{T_2} AR_t \,. \tag{2.1}$$

Next, index the individual CAR values of the F event instances by $f = 1, \ldots, F$, so that one observes a sample of CAR values given by $\{CAR_f\}_{f=1}^{F}$. The cumulative average abnormal return (CAAR), which equivalently can be considered an average cumulative return, is given by

$$CAAR := \frac{1}{F} \sum_{f=1}^{F} CAR_f \,. \tag{2.2}$$

One then is interested in testing

$$H_0 : \mathbb{E}(CAAR) = 0 \quad \text{vs.} \quad H_1 : \mathbb{E}(CAAR) > 0 \tag{2.3}$$

or

$$H_0 : \mathbb{E}(CAAR) = 0 \quad \text{vs.} \quad H_1 : \mathbb{E}(CAAR) < 0 \tag{2.4}$$

or

$$H_0 : \mathbb{E}(CAAR) = 0 \quad \text{vs.} \quad H_1 : \mathbb{E}(CAAR) \neq 0 \,. \tag{2.5}$$

The first two testing problems are one-sided whereas the last one is two-sided; the choice of the particular testing problem is up to the user.

We note that testing CAAR subsumes testing AAR as the special case $m = 1$. Hence, it suffices to develop a method for testing CAAR with arbitrary event-window size m, which is precisely the aim of this paper.

3 Classic test statistic and alternative test statistic

In what follows, we assume that all event instances share a common event-window size m and, for simplicity, a common estimation-window size n.[7] For event instance f, event date and abnormal returns carry a second subscript; for example, the abnormal return of event instance f at time t is denoted $AR_{t,f}$.

The classic test statistic for testing problems (2.3)–(2.5) is

$$t^{\mathrm{c}}_{CAAR} := \sqrt{F}\frac{CAAR}{S_{CAR}} \quad \text{with} \quad S^2_{CAR} := \frac{1}{F-1}\sum_{f=1}^{F}\big(CAR_f - CAAR\big)^2 , \tag{3.1}$$

where the distribution of t^{c}_{CAAR} under H_0 is assumed to follow t_{F-1}, that is, a t-distribution with $F-1$ degrees of freedom. This assumption requires that the $\{CAR_f\}_{f=1}^{F}$ form an i.i.d. sample from a normal distribution with mean zero. In practice, however, such reliance on normality is unsafe, especially since financial abnormal returns are well known to exhibit heavy tails. For large F (with a common rule of thumb being $F \geq 30$), the assumption of normality can be relaxed thanks to the central limit theorem, but this is irrelevant here, as we are concerned with small F, even as small as $F = 2$. Consequently, the classic test cannot be regarded as valid, where "validity" refers to control of the null rejection probability; see Nguyen and Wolf (2024, Remark 3.1).

As will be explained in Sect. 4, using a permutation approach to obtain a p-value rather than relying on the t_{F-1} distribution yields a valid test even when F is small.

The validity of a test concerns its properties under the null. Of equal importance, however, are its properties under the alternative, that is, its power. With power considerations in mind, we propose as an alternative test statistic.[8]

$$t^{\mathrm{a}}_{CAAR} := \frac{CAAR}{S_{CAAR}} \quad \text{with} \quad S^2_{CAAR} := \frac{m}{F^2}\sum_{f=1}^{F} s^2_{n,f} , \tag{3.2}$$

where $s^2_{n,f}$ denotes the unbiased sample variance computed from the abnormal returns of event instance f in its estimation window:

$$s^2_{n,f} := \frac{1}{n-K}\sum_{t=T_{0,f}+1}^{T_{0,f}+n} AR^2_{t,f} . \tag{3.3}$$

[7] Whereas the assumption of a common m is crucial, the assumption of a common n is mainly for expositional convenience and can be relaxed.

[8] A special case of this test statistic applicable to $m = 1$, that is, to testing AAR, has already been proposed in Brown and Warner (1980, Equation (A.11)) We thank an anonymous referee for pointing out this fact.

Here, K denotes the number of parameters estimated in the expected-return model used to compute abnormal returns. For instance, $K = 1$ for the constant-mean model, $K = 2$ for the market model or the CAPM, and $K = 4$ for the three-factor Fama–French model (including an intercept). Formula (3.3) implicitly assumes

$$\frac{1}{n} \sum_{t=T_{0,f}+1}^{T_{0,f}+n} AR_{t,f} = 0 \,,$$

which holds for all commonly used expected-return models.

Test statistic (3.2) generalizes statistic (3.1) of Nguyen and Wolf (2024), which applies to the case of a single event instance ($F = 1$), to the case of multiple event instances ($F \geq 2$). It can be expected to yield a more powerful test than the classic statistic (3.1), since the denominator of the latter relies on only F data points. From another perspective, under normality with equal variance, the null distribution of (3.2) is approximately $N(0, 1)$, whereas the null distribution of (3.1) is t_{F-1}. The variance of the t_{F-1} distribution exceeds one, with $\text{Var}(t_{F-1}) = (F-1)/(F-3)$, for $F > 3$, which is strictly larger than one and decreases only slowly toward one as F grows. Hence, relative to the standard normal distribution, the classic statistic suffers from heavier tails and inflated critical values, leading to lower rejection probabilities under the alternative. In this sense, (3.2) is asymptotically more efficient and can be expected to deliver higher power in finite samples; this is indeed the case, as demonstrated in Sect. 5.

4 Permutation test

To save space, we refer readers unfamiliar with the basics of permutation tests to Nguyen and Wolf (2024, Section 4); for consistency, we also adopt some of their notation. The essence of our procedure is to construct a permutation-based null distribution for a chosen test statistic—either (3.1) or (3.2)—by repeatedly recomputing the statistic on suitably permuted versions of the data. Importantly, the permutation scheme is designed to "keep together what belongs together," thereby guarding against cross-sectional correlation in abnormal returns.

For a given event instance, its combined window of length $n + m$ is defined as the union of its estimation window and its event window. It is possible that multiple event instances share the same combined window; this occurs when a single event affects multiple firms on the same calendar date. Let $W \leq F$ denote the number of unique combined windows, and let $p(w)$ denote the number of event instances (that is, firms) associated with combined window w. We assume that the W unique combined windows are disjoint, that is, that they do not overlap.

When several firms share a common combined window, cross-sectional correlation in abnormal returns may arise, that is, correlation across firms on a given calendar day. Such dependence occurs whenever abnormal returns are not pairwise uncorrelated, or equivalently, when the cross-sectional covariance matrix of abnormal returns on a given day is not diagonal. A leading example is the presence of multiple firms from

the same industry—such as automobiles, banking, insurance, or software—within a combined window. Since commonly used expected-return models do not control for industry effects, abnormal returns of firms within the same industry often exhibit positive cross-sectional correlation. Many standard tests for CAAR in the literature, including the classic t-test and the sign test, are not robust to such dependence. To ensure robustness with respect to cross-sectional correlations, our permutation test applies the same permutation within each combined window.

Accordingly, any permuted dataset is constructed as follows: within each combined window, the same permutation is applied to all firms, whereas different permutations are applied across combined windows.

4.1 The proposed test

The combined sample size, common to all event instances by assumption, is $n + m$, For any $f \in \{1, \ldots, F\}$, let $X_{t,f} := AR_{T_{0,f}+t}$, for $t = 1, \ldots, n$, and let $X_{t,f} := AR_{T_{1,f}+t-n}$, for $t = n+1, \ldots, n+m$, so that

$$\begin{aligned}&\{X_{1,f}, \ldots X_{n,f}, X_{n+1,f}, \ldots, X_{n+m,f}\}\\ &= \{AR_{T_{0,f}+1}, \ldots, AR_{T_{0,f}+n}, AR_{T_{1,f}+1}, \ldots, AR_{T_{2,f}}\}\,.\end{aligned}$$

Next, for $w = 1, \ldots, W$, let $r_w := \{r_{1,w}, \ldots, r_{n+m,w}\}$ be a permutation (or re-ordering) of the set of integers $\{1 \ldots, n+m\}$.

Furthermore, $w(f)$ is a mapping from $\{1, \ldots, F\}$ to $\{1, \ldots, W\}$ such that firm f belongs to combined window $w(f)$.

The permutation of the collection $\{X_{t,f}\}$ is then implied as $X^*_{t,w(f)} := X_{t,r_{t,w(f)}}$ which in return defines the corresponding permutation of the abnormal returns as $AR^*_{T_{0,w(f)}+t} := X^*_{t,w(f)}$, for $t = 1, \ldots, n$, and $AR^*_{T_{1,w(f)}+t-n} := X^*_{t,w(f)}$, for $t = n+1, \ldots, n+m$.

In a nutshell, the permutation test, in its ideal version, works as follows: First, set up the test statistic T in a way such that large values indicate the alternative hypothesis, that is,

$$\begin{aligned}T &:= t_{CAAR} \quad \text{for testing problem (2.3)}\,,\\ T &:= -t_{CAAR} \quad \text{for testing problem (2.4)}\,,\ \text{and}\\ T &:= |t_{CAAR}| \quad \text{for testing problem (2.5)}\,.\end{aligned}$$

Here, the underlying test statistic t_{CAAR} can be either t^{c}_{CAAR} of (3.1) or t^{a}_{CAAR} of (3.2).

Second, for any set of permutations $r := \{r_w\}_{w=1}^W$, denote the value of the test statistic T obtained from the permuted data by T^*_r.

Third, compute the p-value as

$$\hat{p} := \frac{\#\{T^*_r \geq T\}}{\#\{T^*_r\}}\,; \tag{4.1}$$

that is, as the proportion of permutation-based test statistics T_r^* that are greater than or equal to the observed test statistic T.

For practically relevant sample sizes, the total number of distinct permutation sets $r := \{r_w\}_{w=1}^W$, given by $\big[(n+m)!\big]^W$, is far too large to exhaust. Hence, a feasible p-value must be based on a manageable number B of permutation sets, selected suitably from this full collection. The feasible p-value is then computed as

$$\hat{p} := \frac{\#\{T_r^* \geq T\}}{B} .$$

It is customary to include the identity permutation among the B sets, ensuring that one case yields $T_r^* = T$. For the remaining $B - 1$ sets, we suggest drawing the permutations r_w independently and uniformly at random from the $(n+m)!$ possible permutations of the set of integers $\{1, \ldots, n+m\}$, for $w = 1, \ldots, W$. In practice, B should be chosen as large as computationally feasible, with $B \geq 10,000$ as a recommended minimum.

For completeness, we can now summarize the permutation-test method of computing a p-value by means of the following algorithm.

Algorithm 4.1 *1. Choose the test statistic T according to the testing problem of interest, (2.3), (2.4), or (2.5), as described just above (4.1).*
2. Set $T_{r_1}^* := T$.
3. For $b = 2, \ldots, B$, draw a set of permutations $r_b := \{r_{b,w}\}_{w=1}^W$ by drawing a permutation $r_{b,w}$ of the set of integers $\{1, \ldots, n+m\}$ at random, independently for $w = 1, \ldots, W$. Then permute the data according to r_b as described above and denote the value of the test statistic computed from the permuted data by $T_{r_b}^*$.
4. Compute the p-value as

$$\hat{p} := \frac{\#\{T_{r_b}^* \geq T\}}{B} . \tag{4.2}$$

By the general results on permutation testing of Lehmann and Romano (2022, Section 17.2.1), the resulting p-value (4.2) is exact in finite samples; that is, for any $0 < \alpha < 1$,

$$\text{Prob}\big(\hat{p} \leq \alpha\big) = \alpha$$

under the following set of assumptions (under the null hypothesis):

(i) In each combined window $w \in \{1, \ldots, W\}$, the dataset of abnormal returns constitutes and i.i.d. sample of size $n+m$ and dimension $p(w)$, where $p(w)$ denotes the number of firms associated with combined window w, with mean (vector) zero and covariance matrix $\Sigma > 0$.
(ii) Any two abnormal returns that belong to different combined windows are independent.

Importantly, Assumption (i) allows for cross-sectional correlations of abnormal returns of firms that belong to the same combined window. Of course, if there is only one firm in a given combined window, the covariance matrix $\Sigma > 0$ reduces to a variance $\sigma^2 > 0$.

4.2 Aggregation-based approaches to cross-sectional dependence

An alternative and well-established approach to dealing with cross-sectional dependence in event studies is to aggregate firm-level abnormal returns into a single portfolio return and to conduct inference on that aggregated series, as originally advocated by Brown and Warner (1980). When multiple firms are affected by the same event on a given calendar date, such aggregation yields valid inference even in the presence of cross-sectional correlations of firm-level abnormal returns.

However, a practical difficulty with the portfolio approach is that there is no unique or canonical way to aggregate abnormal returns across firms. There are several natural choices for aggregation schemes, including equal-weighted averages, value-weighted averages, and simple sums; however, one can also consider more complex weighting schemes, such as beta-weighted schemes, which even allow for negative weights, as proposed in Brown and Warner (1980, Appendix A.5). Although all of these aggregation schemes lead to valid tests under cross-sectional dependence, they generally have different power properties, and none dominates uniformly across data-generating processes. As a result, inference based on aggregated returns may be sensitive to the particular aggregation scheme chosen.

Moreover, the availability of multiple plausible aggregation schemes raises concerns about specification searching (or "data snooping"). In the absence of appropriate adjustments, a researcher could experiment with several aggregation schemes and report only the one producing the smallest p-value. Such data snooping compromises control of the null rejection probability and thus results in invalid inference.

The permutation methodology proposed in this paper avoids the need to aggregate firm-level abnormal returns altogether, while still delivering exact finite-sample inference in the presence of cross-sectional dependence. By preserving the joint dependence structure of abnormal returns within each combined window, the test remains valid without committing to a particular aggregation scheme. At the same time, the permutation framework is fully compatible with aggregation: researchers who prefer to work with portfolio returns may aggregate returns using a clearly justified scheme and then apply the permutation test to the resulting series. In this sense, the proposed approach provides a flexible and robust tool that encompasses both aggregated and non-aggregated event-study analyses.

Remark 4.1 [Aggregation Changes CAAR Values] Aggregating firm-level abnormal returns into portfolio returns (for firms that share a common combined window) generally changes the value of CAAR. As a result, an aggregation-based CAAR in general no longer coincides with the CAAR at the individual-firm level. This distinction matters if CAAR is used not only for inference but also as an estimator of the average effect of an event on individual firms. Depending on the chosen aggregation scheme and on how firms are grouped into combined windows, the resulting CAAR may overstate or understate the average firm-level effect. This consideration provides an additional motivation for inference procedures, such as the one proposed in this paper, that operate directly on firm-level abnormal returns without requiring aggregation. □

Table 1 Empirical powers

	tCAAR	$\sigma_2^2 = 1$	$\sigma_2^2 = 2$	$\sigma_2^2 = 4$
$\mu = (1, 1)'$	(3.1)	0.10	0.08	0.08
	(3.2)	0.29	0.21	0.15
$\mu = (2, 2)'$	(3.1)	0.18	0.15	0.14
	(3.2)	0.80	0.63	0.43
$\mu = (3, 3)'$	(3.1)	0.27	0.23	0.21
	(3.2)	0.99	0.93	0.76
$\mu = (1, 2)'$	(3.1)	0.10	0.09	0.08
	(3.2)	0.56	0.41	0.26
$\mu = (2, 4)'$	(3.1)	0.11	0.11	0.11
	(3.2)	0.99	0.93	0.76
$\mu = (3, 6)'$	(3.1)	0.05	0.08	0.11
	(3.2)	1.00	1.00	0.98

This table compares empirical powers of the permutation test based on two difference choices of the underlying test statistic, (3.1) or (3.2), in various scenarios. All empirical powers are based on 50,000 Monte Carlo repetitions

5 Monte Carlo studies

5.1 A brief power comparison

Under assumptions (i)–(ii) above, the permutation test yields an exact finite-sample p-value, regardless of whether t_{CAAR} is defined as in (3.1) or (3.2). Thus, when the null hypothesis holds, the choice of test statistic has no impact on the validity of the test. The decision between the two statistics is therefore guided solely by power considerations, that is, by their behavior under the alternative. As argued earlier, t^{a}_{CAAR} (3.2) can be expected to deliver higher power in general. To examine this issue more formally, we conduct a small Monte Carlo study. Specifically, we consider the case of $F = 2$ event instances with non-overlapping combined windows, a common estimation-window size of $n = 120$, and a common event-window size of $m = 1$.

Using the conventions of Sect. 4.1, we let $\{X_{1,1}, \ldots, X_{120,1}\}$ be an i.i.d. sample with distribution $N(0, \sigma_1^2)$ and let $X_{121,1}$ be an independent observation with distribution $N(\mu_1, \sigma_1^2)$. Furthermore, we let $\{X_{1,2}, \ldots, X_{120,2}\}$ be an i.i.d. sample with distribution $N(0, \sigma_2^2)$ and let $X_{121,2}$ be an independent observation with distribution $N(\mu_2, \sigma_2^2)$. The two combined samples $\{X_{i,1}\}_{i=1}^{121}$ and $\{X_{i,2}\}_{i=1}^{121}$ are independent of each other, which corresponds to a setting of two disjoint combined windows with one firm in each window, that is, $F = W = 2$. We let $\mu_1 \in \{1, 2, 3\}$ and set $\mu_2 = \mu_1$ or $\mu_2 = 2\mu_1$. Furthermore, we let $\sigma_1^2 = 1$ and $\sigma_2^2 \in \{1, 2, 4\}$.

The nominal significance level of the test is $\alpha = 0.05$. All empirical powers are based on 50,000 Monte Carlo repetitions. The permutation test is based on $B = 1,000$ sets of permutations.

Table 1 presents the results which clearly demonstrate that test statistic t^{a}_{CAAR} of (3.2) results in uniformly higher power compared to test statistic t^{c}_{CAAR} of (3.1).

Therefore, our clear recommendation is to use test statistic (3.2) for the permutation test.

5.2 Robustness against cross-sectional correlations

As we have stated before, our permutation test guards against cross-sectional correlations of abnormal returns by applying the same permutation to all firms that share a combined window, without the need for aggregation of firm-level results into a single portfolio return.

If instead a separate permutation is applied to any firm, the test may not control the probability of a Type I error. We will now illustrate this property by a small Monte Carlo study. Specifically, we consider $F \in \{2, 5, 10\}$ firms that share a combined window so that $W = 1$. Again, we use $n = 120$ and $m = 1$.

We will study test properties under the null. To this end, consider an F-variate i.i.d. sample of size $n+m = 121$ from a multivariate normal distribution with mean (vector) zero and covariance matrix Σ that has a common diagonal element 1 and a common off-diagonal element $\rho \in \{0, 0.1, 0.25, 0.5\}$. Hence, ρ represents the (common) cross-sectional correlation of abnormal returns. We restrict attention to non-negative values of ρ, since this is the leading case of interest for abnormal returns in event studies. As explained above, it stands to reasons that abnormal returns of firms that belong to the same industry exhibit positive cross-sectional correlations.

We consider two versions of the permutation test: The robust version we propose in Sect. 4.1, which applies the same permutation to all F firms (Perm-R), and a non-robust version that applies F separate permutations, one per firm, chosen independently of each other (Perm-NR). This means that in step 3. of Algorithm 4.1, instead of drawing a set of permutations $\{r_{b,w}\}_{w=1}^{W}$ independently from all permutations of the set of integers $\{1, \ldots, n+m\}$ one draws a set of permutation $\{r_{b,f}\}_{f=1}^{F}$ independently from all permutations of the set of integers $\{1, \ldots, n+m\}$, and then applies permutation $r_{b,f}$ to firm f. As a consequence, the method always uses different permutations for different firms, even if they share a combined window ("do not keep together what belongs together").

The nominal significance level is $\alpha = 0.05$. All empirical null rejection probabilities are based on 50,000 Monte Carlo repetitions. The permutation test based on $B = 1,000$ sets of permutations.

Table 2 presents the results which clearly demonstrate that the non-robust version of the test fails to control the null rejection probability in the case of cross-sectional correlations of abnormal returns. As expected, the problem becomes worse as the cross-sectional correlation (ρ) increases and also as the number of firms (F) increases.

6 Two real-life examples

6.1 Sudden corporate scandals

We examine the impact of sudden corporate scandals, which were unexpected to non-insiders. Two event dates are considered, each associated with a single firm. The

Table 2 Empirical null rejection probabilities

	Test	$\rho = 0.00$	$\rho = 0.10$	$\rho = 0.25$	$\rho = 0.50$
$F = 2$	Perm-R	0.05	0.05	0.05	0.05
	Perm-NR	0.05	0.06	0.08	0.11
$F = 5$	Perm-R	0.05	0.05	0.05	0.05
	Perm-NR	0.05	0.10	0.17	0.26
$F = 10$	Perm-R	0.05	0.05	0.05	0.05
	Perm-NR	0.05	0.16	0.27	0.40

This table compares empirical null rejection probabilities of the permutation test based on the underlying test statistic (3.2) in various scenarios. The version Perm-R, which is the proposal of Sect. 4.1, using a common permutation within combined windows, whereas the version Perm-NR always uses different permutations for different firms. All empirical null rejection probabilities are based on 50,000 Monte Carlo repetitions

Table 3 Test statistics (3.2) and corresponding two-sided p-values for the null hypothesis $H_0 : \mathbb{E}(CAAR) = 0$ for various event-window sizes m

t^{a}_{CAAR}	−15.44	−16.63	−19.54	−13.09
m	m = 1	m = 3	m = 5	m = 7
p-value	0.002	0.000	0.000	0.000

first is 18 September 2015, when disclosure of the diesel-emissions scandal struck Volkswagen. The second is 18 June 2020, when the announcement of accounting irregularities affected Wirecard. Both firms are listed on the Frankfurt Stock Exchange. Abnormal returns are computed using the market model with the DAX index serving as the market proxy. This is a setting with $F = 2$ event instances and $W = 2$ distinct combined windows.

Following standard practice, we consider symmetric event windows, extending equally to the left and right of the event date. The estimation window spans $n = 120$ trading days and ends 11 trading days prior to the event date. In principle, this setup would allow for event windows of up to 21 trading days. For clarity and comparability, however, we restrict attention to window sizes of $m \in \{1, 3, 5, 7\}$ trading days.

We carry out a two-sided test of the sort (2.5) based on test statistic t^{a}_{CAAR} of (3.2). The number of permutations used is $B = 100,000$. Table 3 presents the results. It can be seen that for all event-window sizes m, there is strong statistical evidence against the null hypothesis in favor of a negative effect (p-value < 0.01).

6.2 Sudden announcements of EU antitrust fines

We examine the impact of sudden announcements of European Union (EU) antitrust fines, which were unexpected to non-insiders. Two event dates are considered, with five and four affected firms, respectively. The first event date is 19 March 2024, when the EU imposed fines on several firms for their participation in an automotive-parts cartel. For each of the firms considered, we specify below the stock exchange on which it is listed and the market proxy used to compute abnormal returns under the market model.

Table 4 Test statistics (3.2) and corresponding two-sided p-values for the null hypothesis $H_0 : \mathbb{E}(CAAR) = 0$ for various event-window sizes m

t^{a}_{CAAR}	−0.54	−1.12	−2.32	−2.25
m	m = 1	m = 3	m = 5	m = 7
p-value	0.637	0.346	0.064	0.072

- SKF Group, NASDAQ Stockholm Exchange, OMXS 30
- JTEKT Corporation, Tokyo Stock Exchange, Nikkei 225
- NSK Ltd.,Tokyo Stock Exchange, Nikkei 225
- NTN Corporation, Tokyo Stock Exchange, Nikkei 225
- NFC Ltd.,Tokyo Stock Exchange, Nikkei 225

The second event date is 16 May 2019, when the EU imposed fines on several firms for their participation in a foreign-exchange spot-trading cartel (the so-called Forex – Three-Way Banana Split cartel). For each of the affected firms, we provide below the stock exchange on which it is listed and the market proxy used to compute abnormal returns under the market model.

- Barclays, London Stock Exchange, FTSE 100
- Royal Bank of Scotland, London Stock Exchange, FTSE 100
- Citigroup, New York Stock Exchange, S&P 500
- JPMorganChase, New York Stock Exchange, DJI 30

This is a setting with $F = 9$ event instances and $W = 2$ distinct combined windows; there are five firms in window $w = 1$ and four firms in window $w = 2$.

Following standard practice, we consider symmetric event windows, extending equally to the left and right of the event date. The estimation window spans $n = 120$ trading days and ends 11 trading days prior to the event date. In principle, this setup would allow for event windows of up to 21 trading days. For clarity and comparability, however, we restrict attention to window sizes of $m \in \{1, 3, 5, 7\}$ trading days.

We carry out a two-sided test of the sort (2.5) based on test statistic t^{a}_{CAAR} of (3.2). The number of permutations used is $B = 100,000$. Table 4 presents the results. It can be seen that for event-window sizes $m = 1, 3$, there is no statistical evidence against the null hypothesis in favor of negative effect (p-value > 0.1), whereas for event-window sizes $m = 5, 7$, there is mild statistical evidence against the null hypothesis in favor of a negative effect (p-value < 0.1).

This example is also useful to highlight how our permutation test guards against cross-sectional correlations in abnormal returns by applying the same permutation to all firms within a shared combined window ("keep together what belongs together"). Indeed, for every $b = 2, \ldots, B$, Algorithm 4.1 only draws two independent permutations of the set of integers $\{1, \ldots, n + m\}$; the first one is applied to all five firms in combined window $w = 1$ and the second one is applied to all four firms in combined window $w = 2$. If instead one draws nine independent permutations[9], one for each firm, the four p-values in Table 4 change to 0.566, 0.253, 0.022, and 0.026, respectively. Therefore, by failing to guard against cross-sectional correlations of abnormal

[9] This approach corresponds to the non-robust Perm-NR version of the permutation test studied in Sect. 5.2.

returns, the evidence against the null hypothesis claimed by the test is overstated relative to reality. (The average cross-sectional correlations of abnormal returns are 0.093 in estimation window $w = 1$ respectively 0.299 in estimation window $w = 2$.)

7 Conclusion

This paper addresses the problem of testing CAR and CAAR in event studies when the number of event instances is small—down to as few as two. Standard tests commonly used in the literature are not meaningful or reliable in such settings. To overcome this limitation, we have proposed a permutation test based on an alternative test statistic. The procedure is nonparametric and robust to cross-sectional correlations of abnormal returns which may arise when some firms share the same event date (and thus the same estimation window and the same event window). A Monte Carlo study has highlighted (i) the advantages of the proposed test statistic in terms of power and (ii) the ability of the permutation test to guard against cross-sectional correlations of abnormal returns. Furthermore, applications to two real-life datasets have demonstrated the practical usefulness of the permutation test.

References

Brown SJ, Warner JB (1980) Measuring security price performance. J Financ Econ 8(3):205–258
Campbell J, Lo A, MacKinlay C (1997) The Econometrics of Financial Markets. Princeton University Press, Princeton, New Jersey
Cowan AR (1992) Nonparametric event study tests. Rev Quant Financ Acc 2:343–358
Kliger D, Gurevich G (2014) Event Studies for Financial Research. Palgrave Macmillan, New York
Kolari JW, Pynnonen S (2011) Nonparametric rank tests for event studies. J Empir Financ 18(5):953–971
Kothari S, Warner J (2007) Econometrics of event studies. In: Eckbo BE (ed) Handbook of Empirical Corporate Finance: Empirical Corporate Finance, vol 1. Elsevier, Amsterdam, pp 3–36
Lehmann EL, Romano JP (2022) Testing Statistical Hypotheses. Springer, New York, fourth edition
MacKinlay AC (1997) Event studies in economics and finance. Journal of Economic Literature 35(1):13–39
Nguyen PA, Wolf M (2024) Single-firm inference in event studies via the permutation test. Empirical Economics 66:2435–2450
Wilcoxon F (1945) Individual comparison by ranking methods. Biometrics Bulletin 1(6):80–83