Evaluating VPIN as a trigger for single-stock circuit breakers

David Abad , Magdalena Massot , Roberto Pascual

PII: S0378-4266(17)30196-6
DOI: 10.1016/j.jbankfin.2017.08.009
Reference: JBF 5192

To appear in: *Journal of Banking and Finance*

Received date: 22 July 2015
Revised date: 6 July 2017
Accepted date: 12 August 2017

Please cite this article as: David Abad , Magdalena Massot , Roberto Pascual , Evaluating VPIN as a trigger for single-stock circuit breakers, *Journal of Banking and Finance* (2017), doi: 10.1016/j.jbankfin.2017.08.009

Highlights

- VPIN rarely signals abnormal illiquidity.
- VPIN only occasionally anticipates price changes leading to actual trading halts.
- The capacity of VPIN to anticipate truly toxic events is limited.
- VPIN limits cannot substitute traditional price limits.
- VPIN-based circuit breakers can be costly in terms of unnecessary trading cessations.

Evaluating VPIN as a trigger for single-stock circuit breakers

David Abad
University of Alicante, Spain
goliat@ua.es

Magdalena Massot
University of the Balearic Islands, Spain
m.massot@uib.es

Roberto Pascual[1]
University of the Balearic Islands, Spain
rpascual@uib.es

This version
July 2017

Acknowledgements

We thank two anonymous referees, Geert Bekaert (Managing Editor), Yakov Amihud, Bidisha Chakrabarty, Terrence Hendershott, Isabel Figuerola-Ferretti, Carolina Manzano, José Penalva, Dominik Roesch, Gonzalo Rubio, Andriy Shkilko, and José Yagüe for insightful comments on our results. We thank audiences at the 52nd EFA Annual Meeting (Baltimore, 2016), Auckland Finance Meeting (Auckland, 2015), 14th EEFS Annual Conference (Brussels, 2015), XXIII Finance Forum (Madrid, 2015), and seminars at Universidad Carlos III (Madrid, 2015), Universidad Cardenal Herrera (Alicante, 2015), and Universitat Rovira i Virgili (Tarragona, 2015). Any errors are entirely our own.

[1] Contact author: Carretera de Valldemossa km. 7.5, 07122 Palma, Mallorca, Balearic Islands, Spain. Phone: +34-971171329

Funding: This work was supported by the Spanish DGICYT projects ECO2010-18567, ECO2011-29751, ECO2013-4409-P, and ECO2014-58434-P. Roberto Pascual also acknowledges the financial support of *Fundación* BBVA.

Abstract

We study if VPIN (Easley, López de Prado, and O'Hara, 2012, *Review of Financial Studies* 25, 1457-1493) is an efficient advance indicator of toxicity-induced liquidity crises and related sharp price movements. We find that high VPIN readings rarely signal abnormal illiquidity, and very occasionally anticipate large intraday price changes leading to actual trading halts. We find significant differences in illiquidity and price impact between VPIN-identified toxic and non-toxic halts, but they tend to vanish when we control for *ex ante* realized volatility. We conclude that the capacity of VPIN to anticipate truly toxic events is limited.

JEL Classification: G10

Keywords: VPIN; BVC; circuit breakers; trading halts; price limits; order flow toxicity.

1. Introduction

In a series of related papers, Easley, López de Prado, and O'Hara (hereafter ELO) introduce a metric of order flow toxicity called Volume-Synchronized Probability of Informed Trading (VPIN).[2] Order flow is regarded as toxic when it adversely selects liquidity providers. VPIN is expected to increase when information events induce unbalanced and accelerated trades over relatively short intervals. High VPIN readings are presumed to concur with or precede illiquidity shortfalls; subsequently short-term volatility peaks.

Using ultra-high frequency data on future contracts, ELO (2011a) find that VPIN achieved abnormally high values in the hours preceding the Flash Crash of May 6, 2010. Also, ELO (2012a) find that short-term volatility is often substantial within VPIN-identified toxic periods. Encouraged by these promising findings, ELO (2012a) recommend that market authorities use VPIN to monitor markets in real time, and eventually trigger a trading halt whenever VPIN signals that liquidity provision is at risk. Circuit breakers often use price limits as triggers.[3] In a recent report (see UKGOS, 2012), the UK Government Office of Science advocates for new forward-looking types of circuit breakers that would use early warning signals as triggers. ELO (2012a) claim that circuit breaker programs triggered by VPIN limits could have this preventive capacity.

We evaluate ELO's proposal by addressing the core question of whether VPIN is a valid proxy and/or a reliable leading indicator for order flow toxicity. Failing to document a strong link between high VPIN readings and both contemporaneous and ex-post liquidity withdrawals would call into question the use of VPIN limits to trigger circuit breakers. Previous studies focus on the lead-lag relationship between VPIN and ex-post short-term

[2] See ELO (2011a, 2011b, 2012a, 2012b, 2014, 2015, and 2016).
[3] An example is the "limit-up, limit-down" (LULD) approved as a response to the Flash Crash (see SEC release 34-67091, May 31, 2012). In Europe, similar systems have been in place for a long time now (see Abad and Pascual, 2010, and Zimmermann, 2013).

volatility.[4] However, ELO (2012a) themselves note that not all volatility is due to toxicity. So, volatility-based tests do not provide a satisfactory answer to the question of whether VPIN does signal toxic order flow. Instead, we evaluate VPIN using objective measures of liquidity and adverse selection costs.

We contribute to the literature in several ways. First, we analyze trading activity, liquidity, and realized volatility around VPIN-limit violations using close-in-time VPIN-identified non-toxic days as a benchmark. We show that VPIN limits are triggered by sudden rises in trading volume. Consistent with previous studies, we also find that VPIN-limit hits often precede extraordinary increases in realized volatility. However, relative spreads after VPIN-limit hits are rarely out of the ordinary, and there are no significant changes in the limit order book depth. So, VPIN-identified toxic events do not appear that toxic when judged by objective measures.

Second, we test if VPIN-limit violations anticipate truly toxic events. Earlier studies examine whether VPIN succeeds in signaling extraordinary events, like the Flash Crash (e.g., ELO, 2012a) or single-stock mini-flash crashes (e.g., Pöppe et al., 2016). Despite the sharp price disruptions associated with these crashes, case studies of unique events do not suffice to judge the effectiveness of any metric as an early-warning signal. Instead, we test VPIN's anticipatory capacity by means of a large sample of recurrent, potentially toxic, single-stock events triggered by unusually large intraday price changes. Namely, our exercise involves 6,740 trading halts for 45 stocks listed in an electronic order-driven market over 12 years.

We start by analyzing how often VPIN-signaled toxic events comprise actual trading halts (hereafter, toxic halts). We find that toxic halts represent only 4.5% to 9.9% of all halts, depending on the VPIN parameterization. The immediate regulatory implication of this

[4] See ELO (2012a), Andersen and Bondarenko (2014), Low, Li, and Marsh (2016), Pöppe, Moos, and Schiereck (2016), Song, Wu, and Simon (2014), Wei, Gerace, and Frino (2013), and Wu et al. (2013).

finding is that VPIN-limits can at most complement, but never substitute, price limits. Moreover, during toxic periods with no price-limit hits, the largest intraday price movements observed are remarkably lower than the minimum intraday price variations market authorities are willing to tolerate.

Next, we examine if VPIN-identified toxic halts are truly toxic. We find that the abnormal realized spread (depth) tends to be higher (lower) around toxic-halts than around non-toxic halts. Moreover, the average price impact of trades, a common proxy for adverse selection costs, shows higher abnormal values around toxic halts than around non-toxic halts. Yet, when we control for *ex ante* differences in volatility, most of the anticipatory capacity of VPIN vanishes. In general, our analysis renders limited support for VPIN's preventive capacity.

Our study reveals that VPIN limits would miss most of the intraday price movements that ultimately lead to price-limit hits. Being imperfect substitutes, we recommend market authorities not to replace price limits with VPIN limits. Moreover, authorities must be aware that circuit breakers triggered by VPIN limits could have a high false positive rate in foreseeing truly toxic illiquidity-driven volatility peaks. Since they would not know *ex ante* which VPIN-limit hits are due to toxicity, executing a VPIN-based circuit breaker could impose a high cost in terms of disrupting the normal functioning of markets with no apparent benefit.

VPIN has been the subject of a heated academic debate. Opponents have questioned every finding in ELO's papers. Above them all, Andersen and Bondarenko (2014, 2015) find that VPIN has no incremental predictive power on short-term volatility beyond trading intensity and volatility itself. The bulk-volume classification (BVC) scheme, introduced by ELO (2012a) to estimate order imbalances within the VPIN approach, is also the object of severe criticism. While proponents (e.g., ELO, 2016) show that BVC is as reliable as any standard

tick-based algorithm (TBA) can be, independent studies (e.g., Chakrabarty, Pascual, and Shkilko, 2015, Andersen and Bondarenko, 2015, and Pöppe et al., 2016) show that TBAs outperform BVC. Despite the controversy, VPIN has already been used in several studies assuming is a solid indicator for order-flow toxicity.[5] Our findings question the reliability of the metric.

This study is related to Andersen and Bondarenko (2014, 2015), Chakrabarty et al. (2015), and Pöppe et al. (2016), who also evaluate the VPIN. They focus on the lead-lag relationship between VPIN and volatility. Instead, our emphasis is on the more fundamental question of whether VPIN can be called a metric of order flow toxicity. Our study is also connected to Bhattacharya and Chakrabarti (2014), Cheung, Chou, and Lei (2015), and Phuensane and Williams (2016). Like us, they evaluate VPIN using recurrent events: IPOs, mandatory calls, and episodes of manipulation, respectively. Unlike them, we do not presume the events of interest are toxic. Instead, we test whether VPIN-signaled toxic halts really differ from non-toxic halts.

Last but not least, our analysis is based on high-frequency quote and trade data from the electronic trading platform of the Spanish Stock Exchange (SSE). The SSE has some features that make it ideal for our study. It is the least fragmented of the major EU stock exchanges. Up to January 2013, the SSE market share on the Spanish large caps was above 90%. On December 2013, it fell to 84%. Therefore, our database contains a large majority of all the trades for the SSE-listed stocks. In highly fragmented markets such as the US, it is a challenge to compute VPIN with data from a single trading platform (e.g., Chakrabarty et al., 2015). Consolidated files also suffer from integrity problems (e.g., Holden and Jacobsen, 2014; O'Hara, 2015; Upson, Johnson, and McInish, 2015). Moreover, our sample spans twelve years and thousands of trading halts. A similar task centered on the recent US single-

[5] See Bhattacharya and Chakrabarti (2014), Borochin and Rush (2016), ELO (2015), McInish et al. (2014), and Van Ness, Van Ness, and Yildiz (2016).

stock circuit breaker program, for example, would be nowadays unfeasible (e.g., Cui and Gozluklu, 2016).

The rest of the paper is structured as follows. In section 2, we provide methodological details. In Section 3, we describe the database and provide market background. In Section 4, we characterize the VPIN-signaled toxic periods. In section 5, we study liquidity and volatility around VPIN-limit violations. In Section 6, we examine if VPIN can anticipate truly toxic events. In Section 7, we summarize the robustness tests. Finally, in Section 8 we conclude.

2. Implementing VPIN

VPIN is a moving average of the absolute order imbalance over the most recent *n* volume increments or buckets,

$$VPIN_{i\tau(n)} = \frac{\sum_{\tau=1}^{n} |OI_{i\tau}|}{nV_i}, \quad [1]$$

where the subscript $\tau = \{1,...,n\}$ represents the volume buckets, and $\tau(n)$ is the last bucket; V_i is the size (in shares) of each bucket, and $OI_{i\tau}$ is the order imbalance in the τ-th bucket, that is, the difference between the volume of trades initiated by buyers (V_i^B) and the volume of trades initiated by sellers (V_i^S). VPIN is defined in volume time rather than clock time in an attempt to capture pieces of new information of comparable relevance arriving to the marketplace (ELO, 2012a).

The VPIN level depends on parameter choices. Therefore, it has to be evaluated in relative terms to its own history. Following ELO (2012a), we use the empirical CDF of VPIN to convert each VPIN reading into a cumulated probability (hereafter, relative VPIN). A toxic event starts when the relative VPIN reaches a critical level or VPIN limit (p). As in Andersen

and Bondarenko (2015), we set p = 0.99. A toxic period ends when the relative VPIN crosses up-bottom a second threshold q = 0.85. We choose q low enough to prevent nested short-lived toxic events that essentially constitute the same event.

ELO (2011a, 2012a) fix the number of volume buckets (n) at 50; we also consider n = 25 and 75. As Andersen and Bondarenko (2015), we compute V_i as a percentage (δ) of the average daily volume over the previous month. In this way, we account for the pronounced rise in daily volume within our sample period. We fix δ at 1/50. According to ELO (2012a), VPIN should be robust to a wide range of choices of n and δ. To generate a bucket, we aggregate consecutive trades till the accumulated volume is V_i. If a trade is for a size greater than needed to fill the bucket, the excess size is assigned to the next bucket.

To estimate the $OI_{i\tau}$ in eq. [1], we consider two alternatives. First, as suggested by ELO (2016), we use the BVC approach. We pre-aggregate volume into bars of a given size. Then, we estimate the buy volume ($\hat{V}_b^B$) within a bar b as

$$\hat{V}_b^B = V_b \times \Phi\left(\frac{\Delta p_b}{\sigma_{\Delta p}}\right), \qquad [2]$$

where V_b is the aggregated volume; $\Delta p_b = p_b - p_{b-1}$ is the price change between two consecutive bars, where p_b is the price of the last trade in bar b; $\sigma_{\Delta p}$ is the volume-weighted standard deviation of Δp_b, and $\Phi(.)$ is the CDF of the probabilistic distribution assumed for $\Delta p_b / \sigma_{\Delta p}$. The relative weights of buy and sell volume within a bar depend on how large Δp_b is relative to the assumed distribution of price changes.

BVC-based $OI_{i\tau}$ estimates and, thus, VPIN estimates, depend on the choice of the type and size of bars, and the assumed CDF of $\Delta p_b / \sigma_{\Delta p}$. In our exercise, we use time, trade, and

volume bars. We consider time bars from 30 to 1800 seconds in increments of 30 seconds. Volume and trade bar sizes are stock-specific to account for differences in the average daily trading activity across assets and overtime. Volume (trade) bar sizes are computed as the closest integer to υV_i (υT_i), υ varying between 1% and 40% in increments of 1%. T_i is the average daily number of trades over the preceding month.

Regarding the distribution of $\Delta p_b / \sigma_{\Delta p}$, ELO (2011a) and Andersen and Bondarenko (2014) assume normality. More recently, however, ELO (2016) and Wu et al. (2013) suggest t-student distributions with 0.25 and 1 degree of freedom, respectively. Our main results are independent on this particular choice. Without loss of generality, we provide results with the t-student with 0.25 degrees of freedom.

In the Appendix, we summarize all our parameter choices for the BVC, resulting in 840 different estimates of VPIN and 1680 different sets of toxic events per stock. Hereafter, we will refer to the VPIN with BVC simply as "VPIN".

Second, our database contains and initiator flag. So, we can compute VPIN using the actual initiator-based order imbalance (hereafter, VPIN-flag). Evaluating the performance of VPIN-flag is important for several reasons. On the one hand, there is no need to calibrate unobserved parameters; determine the proper type of bars, or their size; choose an underlying distribution for the changes in prices, etc. The resulting VPIN-flag is unique given the starting point of the series. On the other hand, the trade initiator plays a fundamental role in market microstructure, with literally hundredths of empirical studies relying on it (or estimates of it). Even in the VPIN literature, VPIN-flag has already been used by several papers, including Andersen and Bondarenko (2015), Chakrabarty et al. (2015), Panayides, Shohfi, and Smith (2014), and Pöppe et al. (2016). Recently, ELO (2016) strongly object to initiator-based $OI_{i\tau}$ estimates arguing that the aggressor flag becomes distorted in the

presence of high-frequency trading (HFT). However, in ELO (2012a), they state that "the trade-classification algorithm itself is independent of the VPIN metric" and "any algorithm could be used to provide input to the estimation of VPIN" (p. 1465).

3. Market background and data

We use 12 years (2002-2013) of intraday trade and quote high-frequency data from the electronic trading platform of the SSE, called *Sistema de Interconexión Bursatil Español* (SIBE). According to the World Federation of Exchanges, in 2013 the SSE was the fifth exchange by domestic market capitalization, and the third by total value of share trading in the Europe-Africa-Middle East Region. The SIBE handles the trading activity of the most liquid Spanish stocks. Trading is continuous from 9:00 am to 5:30 pm GMT+1, with regular call auctions at the opening (8:30-9:00 am) and closing (5:30-5:35 pm). We only use data from the continuous trading phase.

Our database comprises limit order book (LOB) and trade files. LOB files include snapshots of the five best ask and bid quotes, updated after each order submission, revision, or cancellation. For each quote, we know the displayed depth, but not the hidden volume due to iceberg orders (e.g., Pardo and Pascual, 2012). Liquidity supply depends entirely on the LOB since there are not designated market makers. For each trade, we know its price, its direction (i.e., buy or sell), its size, and the best quotes prevailing before the trade. We only keep ordinary trades. We exclude trade registers containing the allocation price and volume of the corresponding auction. Time stamps in both files are in hundredths of a second (milliseconds after June 2013). Trade and LOB files can be perfectly matched using common sequence code.

We consider stocks that traded continuously during at least three years within our sample period and were constituents of the IBEX-35, the official market index, at least half of the

time. We drop temporarily delisted stocks or stocks with prices below one euro. Our final sample consists of 45 stocks. In Table I, Panel A, we provide cross-sectional average daily statistics for the whole sample and the ten largest and ten smallest stocks. In Table 1, Panel B, we provide cross-sectional average statistics on the evolution of trading activity and liquidity over the sample period.

[Table 1]

Trades (volume) grew by 257% (62.9%) between 2002 and 2013. Despite the short-sale ban in place from July 23, 2012, to January 31, 2013, most of that growth happened post-MiFID (2008-2013). Trade size halved, mostly post-MiFID. Message traffic (i.e., order submissions, revisions, and cancellations) experienced an extraordinary 1,958% increase, suggesting more intense HFT activity (e.g., Angel, Harris, and Spatt, 2011). Message traffic per trade, a common proxy for HFT (e.g., Hendershott, Jones, and Menkveld, 2011), raised 323.7% post-MiFID.

Regarding liquidity, the relative spread halved over the sample period. As it happened worldwide (e.g., Beber and Pagano, 2013), SSE relative spreads increased during the recent financial crisis, but they decreased 14.81% over the post-MiFID period. LOB depth raised 273.27% during the pre-MiFID period but decreased 300% post-MiFID. A staged program initiated in May 2009 to reduce the tick size from 0.01 to 0.001 or 0.005, depending on the stock price, could partly explain this structural change (e.g., Goldstein and Kavajecz, 2000). Reduced tick sizes cheapened competition for price priority, tightening the book near the best quotes (see "dispersion" in Table 1). We control for tick size changes in posterior analyses.

Since May 2001, the SSE implements a single-stock circuit-breaking mechanism to handle episodes of extraordinary volatility. It consists of "static" and "dynamic" price limits that trigger short lived call ("volatility") auctions. Static (dynamic) price limits set the

maximum permitted variation around the allocation price of the last auction (the last trade price). Price ranges are revised every month based on the stock's volatility over the last six months. Non-regular revisions may also occur.

A volatility auction starts when an incoming order is about to execute at a price at or above (below) the upper (lower) price limit. It lasts five minutes plus a random end of at most thirty seconds to avoid price manipulation. Never mechanically extended, discretionary extensions may happen. We find 50 cases, lasting about 27 minutes on average. They do not drive our results. After the auction, new price limits are set around the allocation price. While a violation of the static price limit implies a remarkable intraday variation in the stock price, a violation of the dynamic price limit may happen because a single incoming aggressive order encounters an unusually thin book. Within our sample, there are 6,740 price limit violations; 3,794 (56.3%) dynamic, and 2,946 (43.7%) static. During ordinary times, static ranges vary from 4% to 8%, but ranges up to 20% have been imposed under extreme market turmoil, like October 2008. Dynamic ranges fluctuate between 1% and 10%. For a given stock, the dynamic range is always smaller than or equal to the static range.

4. Toxic events

In Table 2, we provide cross-sectional summary statistics on the number and persistence of VPIN-identified toxic events. For each bar type, we provide cross-sectional medians across different bar sizes. VPIN signals 30 to 46 toxic events per stock. The median toxic event comprises between 55 and 78 buckets. Using time bars, toxic events persist about five hours in average, two hours less than using trade or volume bars. As a result, intraday toxic events are less common with volume/trade bars than with time bars.[6] VPIN-flag identifies 7 toxic events per stock of much longer average duration (447 buckets). The lengthy duration of

[6] The accuracy of BVC depends on the average number of trades within a bar (Chakrabarty et al., 2015) and the variability in the number of trades across bars (ELO, 2016). Differences in accuracy across bar types could explain the reported differences in persistence of toxic events.

toxic events might be of some concern. Being a moving average, VPIN might keep signaling high toxicity after a VPIN-triggered halt even when the halt succeeds in restoring regular trading.

[Table 2]

We wonder if different VPIN parameterizations result in the same VPIN-limit violations. As the VPIN performance must ultimately depend on the accuracy of the order flow imbalance estimates, our focus is on the BVC parameters. In Panel A of Table 3, we change the bar size within each bar type. In Panel B of Table 3, we vary the bar type. In each case, we randomly pick two different VPIN specifications and obtain the percentage of overlapped events. We repeat this process 200 times, and compute median statistics across random pairs.

[Table 3]

From Table 3, we learn that the VPIN performance as a trigger is contingent upon BVC parameter choices. In Panel A, we show that the percentage of overlapping toxic events across bar sizes varies between 58.35% for volume bars and 70.72% for time bars. VPIN specifications disagree more when we randomly pick distant bars. For example, the percentage of overlapped toxic events for trade bars is 77.3% (52.8%) when the difference between bar sizes is less (more) than 10% (20%) of the average daily number of trades. In Panel B, we show that VPIN is highly sensitive to changes in the type of bar. For example, only 29% (47%) of the toxic events identified using time bars are found to be toxic using volume (trade) bars.

Our results agree with previous studies suggesting that VPIN is sensitive to parameter changes (e.g., Andersen and Bondarenko, 2015; Chakrabarty et al., 2015), and disagree with

ELO's (2012a) claim that the bar size must have a minor influence on the value of VPIN.[7] Recognizing that the performance of price limits is also contingent on choices such as the reference price or the allowed price range, our findings thus far do not necessarily undermine VPIN as a trigger. However, they stress the need to carefully tune the metric. We will address the optimal calibration of VPIN in the robustness section. For the rest of the paper, we will use uniform bar types and sizes for all stocks.

5. VPIN as a proxy for order flow toxicity

According to ELO (2012a), VPIN-limit violations signal the highest relative levels of order flow toxicity within our sample period. According to the adverse selection costs literature (e.g., O'Hara, 1995; Foucault, Pagano, and Röell, 2013), VPIN-limit hits should precede extraordinary liquidity withdrawals and, as a result, peaks in short-term volatility. In this section, we evaluate the reliability of VPIN as a proxy for order flow toxicity by testing the expected connection between VPIN-limit hits and illiquidity.

As an initial test, we examine cross-sectional average abnormal illiquidity levels around VPIN-limit violations. For each VPIN-limit hit, we compute trading activity, liquidity, and realized volatility statistics for 24 five-minute intervals centered on the time of the VPIN-limit violation. We exclude intervals that start after the estimated termination of the toxic event. We compute realized volatility (*RV*) as the standard deviation of the one-minute returns within a given interval. As liquidity metrics, we use the relative spread (*RS*), that is, the quoted bid-ask spread divided by the quote midpoint, the average depth in euros at the market quotes (*DB*), and the average accumulated depth in euros at the five best ask and bid quotes of the LOB (*DK*), all of them averaged weighting by time. For trading activity metrics, we use the volume in shares (*VOL*) and the number of trades (*TRD*). Abnormal market

[7] ELO (2012a) rely on a limited exercise with only time bars, a single asset (the E-min S&P500), and a single event (the Flash Crash).

conditions are evaluated as follows: for each event, we use the closest non-toxic 250 days with the same tick regime as benchmark; then, we standardize each metric by subtracting its mean and dividing by its standard deviation over the benchmark days and during the same time interval.

In Figure 1, we plot the estimated cross-sectional average abnormal levels for VPIN, with selected uniform bar sizes, and VPIN-flag. Our findings are robust across bar sizes. Thus, we provide results for VPIN with 60-second time bars, and volume and trade bars of size υ=2%. In Figure 1.a, we show that VPIN-limit violations happen because of abrupt jumps in volume within the last five minutes (*t*=-1). With time bars, for example, *VOL* deviates from the benchmark mean 16.48 times the benchmark standard deviation (hereafter, STD). Abnormal levels persist shortly after the VPIN-limit violation, but they are never as high as in *t*=-1. *TRD* (not reported) increases much less than *VOL*, revealing that VPIN limits are ultimately hit by an unusual concentration of relatively large trades.

[Figure 1]

In Figure 1.b, we show that high VPIN readings precede short-lived abnormally high volatility realizations (*RV*), consistent with ELO (2012a), but they are also headed by extreme *RV* readings. Namely, *RV* progressively increases before the VPIN-limit is hit, peaks either immediately before (for time and trade bars) or immediately after (for volume bars and VPIN-flag) the hit, and then quickly falls below pre-hit levels. Therefore, Figure 1.b documents a strong contemporaneous correlation between VPIN and realized volatility.

We should expect extraordinary liquidity withdrawals after VPIN-limit violations. Although liquidity declines on average, the change in *RS* reported in Figure 1.c is less dramatic and persistent than expected. For VPIN, *RS* is 0.7 to 1.04 STD above the benchmark mean before the VPIN limit is hit, from *t*=-12 to *t*=-1, which is consistent with the presence

of toxic order flow. Immediately after VPIN-limit violations, *RS* reaches a maximum of 1.06 (0.85) STD for volume (time) bars, to quickly fall below pre-hit levels. For VPIN-flag, we find no remarkable deviation in *RS* from the benchmark levels. Regarding quoted depth, Figure 1.d shows that *DB* before the VPIN-limit hit is only slightly below ordinary levels when we use volume bars. After the hit, it decays, but the drop is either not statistically significant (VPIN) or remains above ordinary levels (VPIN-flag). The patterns we obtain for abnormal *DK* (not reported) are similar.

The low average correlation between high VPIN readings and illiquidity suggests that VPIN often fails as an indicator of order flow toxicity. To gain further insights on this core issue, we take a closer look at liquidity and volatility around each VPIN-limit hit. Per stock-event, we take the same 250 benchmark days as before, split their trading sessions into regular five-minute intervals, compute the standardized liquidity and volatility proxies and obtain the 1^{st}, 5^{th}, 10^{th}, 90^{th}, 95^{th} and 99^{th} percentiles of the resulting empirical distribution. We classify *RS* or *RV* (*DK*) as "extraordinary" if it exceeds (is lower than) a focal RHS (LHS) benchmark percentile. Under the null of ordinary market conditions, *RS* should be found to be extraordinary with respect to the, for example, 95^{th} benchmark percentile in about 5% of the VPIN-signaled toxic events.

In Table 4, we provide the proportion of VPIN-limit hits preceded or followed by extraordinary illiquidity and volatility levels for a 30-minute window centered on each VPIN-limit hit and split into five-minute intervals. We also provide average deviations with respect to the benchmark percentiles across all toxic events.

[Table 4]

We first look at *RV*. Consistent with Figure 1, VPIN-limit violations often precede extremely high *RV* levels. Using 1-minute time bars, for example, VPIN-limit hits are

immediately followed by extreme volatility with respect to the 99th benchmark percentile in 39.1% of the events. Somewhat unexpected is that *RV* is often high before extreme VPIN realizations. With time bars, 49.56% of the events are shortly preceded by extremely high *RV*. Andersen and Bondarenko (2015, p.39) conclude that VPIN has predictive power on future volatility "due to its correlation with realized volatility, which arises from the use of price changes to infer order imbalance" (i.e., BVC). Our findings are consistent with their conclusions. Indeed, the connection between VPIN-flag and *RV* is much weaker than that between VPIN (with BVC) and *RV*, with only 12.89% (9.14%) of the limit hits being followed (preceded) by extraordinarily high realized volatility.

We now turn our attention to our main concern: liquidity. Table 4 reveals that most of the VPIN-limit hits are not toxicity-driven. Using BVC (Panel A to C), we find *RS* in the [0 5) interval to be above the 99th benchmark percentile in 12.2%-14.09% of the toxic events. The occurrence of extreme *RS* readings declines in posterior intervals. The rate of success of VPIN in anticipating liquidity shocks is therefore much lower than that obtained for realized volatility. Special mention deserves VPIN-flag, for which *RS* is rarely above the 99th benchmark percentile. Our findings therefore question the toxic nature of many of the VPIN-signaled volatility peaks.

With respect to the depth dimension of liquidity, captured by *DK*, liquidity following VPIN-limit hits is nothing but ordinary. For VPIN with time bars (Panel A), the rate of success for the [0 5) interval, with respect to the 1st benchmark percentile, is only 3.28%. With volume bars (Panel B) and trade bars (Panel C), post-event *DK* is extraordinarily low in only 4.73-5.01% of the occasions. VPIN-flag (Panel D) renders the lowest rate of success, 2.39%.

To sum up, according to ELO (2011a, 2012a), high VPIN readings signal that order flow toxicity is at its peak within the evaluated period. Under these circumstances, liquidity

provision should be at risk. Contradictorily, our analysis shows that liquidity providers do not withdraw from the market soon after a VPIN-limit hit. The severe short-lived increases in short-term volatility we do observe after VPIN-limit hits rarely follow exceptionally high illiquidity. In other words, VPIN-identified events do not appear that toxic when judged by objective measures. Our findings question the reliability of VPIN as a proxy for order-flow toxicity and cast doubt on the appropriateness of halting the continuous trading session after every VPIN-limit hit.

6. VPIN as an early warning signal for toxic events

We have corroborated that within VPIN-identified periods of high and persistent toxicity volatility is often substantial. Since price limits are hit by unusually large price movements, periods of high and persistent toxicity might also comprise actual trading halts. However, ELO (2012a, 2014) argue that extreme volatility should not always be preceded by high VPIN levels, since not all volatility is due to toxicity. Accordingly, trading halts that do not fall within toxic periods should be of a different nature than those that do fall. In this section, we formally test this hypothesis.

In Figure 2, we plot cross-sectional average abnormal levels of the relative spread (*RS*), volume in shares (*VOL*), and realized volatility (*RV*) around static (Figure 2.a) and dynamic (Figure 2.b) halts. We consider twenty-four five-minute intervals centered on each trading recession. As the benchmark, we take the 250 days closest in time to the event day with no trading halts and the same tick regime. We do not control for toxicity because, as previously shown, the same day could be toxic or non-toxic depending on the particular parameterization of VPIN we choose. We standardize each observation using the benchmark values for the same metric over the same five-minute interval. Figure 2 uncovers the different nature of the two types of SSE halts. Static halts are preceded by progressive increases in *VOL* and *RV*. In

contrast, dynamic halts are the result of liquidity shortfalls that enhance *RV*. For both static and dynamic halts, *RS* reaches its highest point right after the halt. Abnormal market conditions persist at least one hour after the continuous session resumes.

[Figure 2]

We proceed to evaluate if, according to VPIN, the abnormal volatility found around SSE trading halts is due to order flow toxicity. We drop six halts that occur before the corresponding VPIN series is initiated, that is, before we collect the first 50 volume buckets. We are left with 6,734 halts, 2,943 of which are static. We classify a trading halt as "toxic" if it totally or partially falls within the limits of a toxic period. In Table 5 Panel A, we show that the majority of the SSE halts happen in periods of no remarkable toxicity. Depending on the bar type, VPIN classifies between 7.19% and 9.7% of the trading halts as toxic. If we control for the type of halt, VPIN classifies 5.47%-11.69% of the static halts and 6.81%-8.52% of the dynamic halts as toxic. According to VPIN-flag, only 4.44% of the halts are toxic.

[Table 5]

In Panel B of Table 5, for toxic events that comprise at least one trading halt, we provide the average distance from the VPIN-limit hit to the closest price-limit hit. We exclude overnight periods, holidays, and weekends. A VPIN-limit violation precedes the closest toxic halt by about 78 minutes for time bars, more than 3 hours for volume bars, and more than 15 hours for VPIN-flag. As a result, the percentage of toxic halts occurring during the session in which the VPIN limit is reached decreases from 48.82% for VPIN with time bars to 7.7% for VPIN-flag.

We conclude that VPIN limits miss most of the abnormal intraday price movements leading to price-limit hits. Even if we are willing to accept that high VPIN readings could anticipate truly toxic events occurring one or more trading sessions ahead, our results indicate

that only a few of the SSE halts are actually anticipated by VPIN-limit violations. Therefore, VPIN limits can complement but never replace traditional price limits.

A plausible explanation for our finding is that only a few of the SSE trading halts, those identified by VPIN, are actually triggered by order flow toxicity. An alternative explanation is that VPIN-identified toxic periods comprise actual halts just by chance. The rest of this section is directed to discern which of these explanations gets greater support in our data.

In Table 6, we compare standardized liquidity (*RS*, *DK*) and realized volatility (*RV*) around VPIN-identified toxic halts vs. non-toxic halts. We also provide statistics on a commonly used metric of adverse selection costs – the trade-size weighted average price impact of trades (*PI*). The price impact of a trade is computed as the difference between the quote midpoint one minute after the trade and the quote midpoint prevailing before the trade. In this case, we standardize the variables using the 250 days closest in time to the event (i.e., halt) day, with the same tick regime, no trading halts, and no toxicity (i.e., relative VPIN<0.9). We provide results for static halts only.[8] Our focus is on the 15-minute window before the price-limit hit and the 15-minute window after the resumption of the continuous session.

[Table 6]

For VPIN, we find *RV* around toxic halts to be significantly higher than around non-toxic halts. Illiquidity, as measured by *RS*, is also relatively higher around toxic halts. Regarding LOB depth, we find that *DK* is significantly lower around toxic halts, but only with trade or volume bars. These findings suggest that VPIN might occasionally succeed as an early warning signal for truly toxic events. Consistently, *PI* is significantly higher around toxic halts than around non-toxic halts. For VPIN-flag, however, our findings show that this metric

[8] Results for dynamic halts are totally consistent and can be found in the online appendix.

fails as an advanced indicator of toxicity. No significant differences in volatility, liquidity, or price impact emerge between VPIN-flag identified toxic and non-toxic halts.

Although our findings in Table 6 provide some support to VPIN-based circuit breakers, their value added would be questionable if differences in liquidity and *PI* between toxic and non-toxic halts can be explained by differences in *ex ante* volatility (e.g., Andersen and Bondarenko, 2015). We run a regression analysis to evaluate the incremental predictive power of VPIN over *RV*. We estimate the model in eq. [3] by OLS with White-robust standard errors

$$L_{ht} = \alpha + \beta_T Toxic_h + \beta_R Range_h + \sum_{j=1}^{3} \beta_{Vj} RV_{ht-j} + \sum_{y=3}^{13} \beta_Y^y Y_{ht}^y + \sum_{s=2}^{45} \beta_S^i S_{ht}^i + \varepsilon_{ht} \quad [3]$$

where L is the standardized liquidity metric for trading halt h and five-minute interval t; *Toxic* is a dummy variable that equals one for VPIN-identified toxic halts; *Range* is the stock-specific price range in place when the trading halt happens; *RV* is the standardized realized volatility metric; Y^y, for y = 2003 to 2013, are year dummies, and S^i, for i = 2 to 45, are stock dummies. The dependent variable L is, in turn, *RS*, *DK*, or *PI*, defined and standardized as before. The coefficient of interest is β_T, which captures differences in each dependent variable between toxic and non-toxic halts. We use lags rather than the contemporaneous *RV* to avoid potential endogeneity problems. Finally, we also consider a model with lagged volume statistics. Our findings are virtually the same.

In Table 7, we provide the estimated β_T in eq. [3] for static (Panel A) and dynamic (Panel B) halts. For each type of halt, we provide findings for VPIN (with time, volume, and trade bars) and for VPIN-flag. Due to space considerations, we only present results for the five-minute interval preceding the price-limit violation and the five-minute interval after the halt.

[Table 7]

For VPIN with time bars, the differences in liquidity and price impact between toxic and non-toxic static halts reported in Table 6 vanish once we control for differences in *ex ante* volatility. For VPIN with volume or trade bars, *RS* and *PI* are still found to be significantly higher (β_T>0) around toxic static halts than around non-toxic static halts. Differences in LOB depth, however, disappear. Similarly, for dynamic halts, toxic halts detected by VPIN with time or volume bars show higher *RS* and *PI* immediately after the VPIN-limit hit than non-toxic halts. For VPIN with trade bars, differences in *ex ante* realized volatility explain the findings in Table 6. Finally, for both static and dynamic halts, Table 7 corroborates the poor performance of VPIN-flag as an early warning signal for toxic events. In general, based on our findings we cannot reject that a properly calibrated VPIN may occasionally succeed in anticipating truly toxic events. It is important to realize that this preventive ability is, nevertheless, limited, since VPIN has a high false positive rate in signaling order flow toxicity.

7. Robustness analyses

In this section, we summarize extra analyses that provide further support to our main findings. Unreported tables are provided in the online appendix of the paper.

7.1. Robustness of VPIN in signaling toxic halts

In Section 4, we show that VPIN is sensitive to changes in key design parameters when signaling toxic periods. We have also studied the sensitivity of VPIN in signaling toxic halts. We find that only 43% of the toxic halts identified by VPIN with 60-second time bars are classified as toxic by VPIN with v=2% volume bars. The highest matching score, involving time and trade bars, is 53%. The lowest matching scores, between 2.5% and 4.1%, involve VPIN-flag.

7.2. Toxicity around trading halts

In Section 5, we show that VPIN-limit hits often precede extraordinary realized volatility (*RV*), but seldom anticipate extraordinary relative spreads (*RS*). We perform a similar analysis around price-limit hits using the 250 days closest to the event day, with no halts, and the same tick regime as benchmark. We find that both *RS* and, foremost, *RV* are severe around the price-limit hits as often as around VPIN-limit hits. Yet, opponents argue that circuit breakers may exacerbate volatility (e.g., Kyle, 1988; Lehmann, 1989; Madhavan, 1992). Therefore, high *RV* readings around price-limit hits could be caused, at least partially, by the circuit breaker itself.

7.3. Restarting VPIN

VPIN measures toxicity in relative terms to its own history. As a result, the metric might fail to identify toxic periods that are recent highs but not historical highs. Restarting VPIN periodically could alleviate this potential limitation. To provide some insights on this issue, we shift the starting point of our VPIN series to December 2008 (VPIN09). In Table 8, we show that VPIN09 signals more toxic events from 2009 to 2013 than VPIN with starting point January 2002 (VPIN02). Moreover, except for VPIN-flag, VPIN09 locates more toxic halts than VPIN02. Indeed, VPIN09 signals as toxic most of the VPIN02-identified toxic halts; in contrast, VPIN02 misses many toxic halts signaled by VPIN09. Therefore, we confirm that VPIN ignores recent toxic highs unless it is restarted regularly. Determining how often to restart VPIN could represent an extra challenge for market authorities.

[Table 8]

7.4. VPIN performance in high-frequency and low-frequency environments

ELO (2011, 2012a) design the VPIN to be a warning signal of order flow toxicity in high frequency environments. Arguably, HFT should be in an embryonic stage during most of the

first-half of our sample period.[9] However, our main findings still hold if we restrict ourselves to the second half of our sample and use VPIN09 instead of VPIN02. We interpret this as evidence that our conclusions are not driven by the low-frequency part of our sample.

7.5. Extreme price changes within VPIN-identified toxic events

In Section 6, we show that most of the VPIN-identified toxic periods do not comprise actual trading halts. However, SSE market authorities revise static and dynamic price limits periodically according to the most recent historical volatility of the asset. So, it could be the case that toxic periods do contain extraordinary price changes, but the price ranges in place are so wide that price limits are rarely hit. In Table 9, we explore this possibility. We provide cross-sectional average statistics on the maximum dynamic (Panel A) and static (Panel B) price variations within toxic periods with no price-limit hits. No matter the VPIN version, we find that the average maximum (both static and dynamic) price variation observed during VPIN-signaled toxic periods is much lower than the corresponding benchmark statistic. Moreover, for 70 to 80% (81% to 92%) of the toxic events signaled by VPIN, price changes never reach the minimum dynamic (static) range of 1% (4%) that can be assigned to an SSE-listed stock. So, SSE authorities would label the most extreme price variations observed within VPIN-identified toxic periods as "tolerable". This finding is at odds with ELO's (2012a, p. 1486) claim that it takes persistently high levels of VPIN to "reliably generate large absolute returns".

[Table 9]

7.6. Extreme illiquidity within VPIN-identified toxic events

[9] The Spanish *Comisión Nacional del Mercado de Valores* (CNMV) estimate that HFTs account for 25-30% of the SSE volume traded in 2010 (see CNMV, 2011). For twelve large SSE-listed stocks in 2013, the European Securities and Markets Authority (ESMA) attributes 32% of the euro volume traded, 29% of the trades, and 46% of the orders to HFTs (see ESMA, 2014).

In Section 5, we conclude that most of the VPIN-limit hits do not appear to be toxic when judged by objective liquidity metrics. This conclusion is based on an examination of how liquidity behaves immediately before and after VPIN-limit hits. However, if VPIN were an effective advance indicator of order flow toxicity, liquidity withdrawals could happen later in the toxic period. In Table 10, we investigate this possibility. We split each VPIN-identified toxic event with no trading halts into regular five-minute intervals, compute the average relative spread weighted by time (*RS*) per interval, standardized as in previous tests, and pick its maximum realization. We compare this value with the maximum of the standardized *RS* for the same five-minute interval over the corresponding 250 benchmark days. No matter the VPIN parameterization, we find significantly worse liquidity realizations during VPIN-identified non-toxic days than during VPIN-identified toxic events. This analysis reinforces our main conclusion that VPIN is neither a reliable proxy nor an effective advance indicator for order flow toxicity.

[Table 10]

7.7. Calibration

How should market authorities optimally calibrate VPIN? ELO do not address this question. Wu et al. (2013) (hereafter, WGLR) consider 16,000 VPIN parameter combinations for futures contracts and rank them by how often high VPIN readings lead extremely high in-sample realized volatility peaks. WGLR's in-sample optimization approach raises concerns. Should anyone be impressed if, after searching over thousands of VPIN versions, we find a constellation of parameters for which VPIN aligns reasonably well with the spikes of realized volatility? Some would probably consider this approach as ultimate data snooping. Should an in-sample optimized VPIN be optimal out-of-sample too? Probably not; we have already seen that it is convenient to restart VPIN periodically. Finally, not all volatility is due to toxicity. Therefore, WGLR's approach might lead to suboptimal results.

Despite all our concerns, we calibrate VPIN just to show that even when VPIN is helped with an in-sample calibration our conclusions persist. Our calibration exercise differs from WGLR in several ways. We fix all the VPIN parameters except the bar type and size because, as we have already shown, the ultimate performance of VPIN depends on BVC. In WGLR, the BVC parameters are not free. Albeit limited, our calibration exercise suffices to gauge whether VPIN improves by fine-tuning some key parameters. We optimize parameters per stock, whereas WGLR render uniform optimal parameters. Finally, we label abnormal *RV* (*ARV*) as "extreme" when it exceeds the 99th percentile of its empirical distribution over the benchmark days.[10] WGLR are more lenient with VPIN; they take the benchmark average *RV* as threshold. We calibrate VPIN over the five-minute interval following each VPIN-limit violation.

Using the calibrated VPIN (hereafter, CVPIN), we corroborate our main findings. The percentage of CVPIN-limit hits immediately followed by extreme *ARV* is 56.1%, much higher, as expected, than with uniform bars (21.4%-36.4%). However, differences in ex-post liquidity are modest or negligible, meaning that CVPIN also has a high false positive rate in signaling toxic order flow. The majority of toxic periods signaled by CVPIN do not contain actual trading halts. In Table 11, we use CVPIN instead of VPIN to replicate the analysis summarized in Table 7. For CVPIN with time or trade bars, differences in illiquidity (*RS*) and price impact (*PI*) between CVPIN-signaled toxic and non-toxic halts disappear once we control for ex ante volatility. CVPIN with volume bars still shows some preventive capacity, mostly for dynamic halts.

[Table 11]

[10] The abnormal *RV* (*ARV*) is the deviation of *RV* from its benchmark mean divided by its benchmark standard deviation. The benchmark consists of the 250 days closest to the event day with no toxicity (VPIN< 0.9), and the same tick size regime.

8. Conclusions and final comments

Financial markets around the world rely on circuit breakers triggered by price limits to constrict volatility. Several recent voices question the usefulness of these traditional systems in high-frequency markets and advocate for forward-looking types of circuit breakers. ELO (2012a) propose the use of VPIN, a new order flow toxicity metric, as a trigger for such type of circuit breaker. We evaluate ELO's proposal by answering two core questions: Is VPIN reliable as a proxy for order flow toxicity? Is VPIN efficient as an early warning signal for toxic events?

Consistent with previous studies, we find that within periods of persistent high VPIN readings volatility is often substantial. Yet, VPIN-limit hits are also frequently preceded by high volatility realizations, providing support to Andersen and Bondarenko's (2015) critique to VPIN. We contribute to the debate by showing that most of the volatility peaks signaled by VPIN are not toxic when judged by objective liquidity metrics. In other words, VPIN is not reliable as a proxy for order flow toxicity. Thus, implementing VPIN-based circuit breakers might result in elevated costs in the form of unnecessary trading cessations. Regulators and market authorities should take note of this flaw of the VPIN approach.

When confronted with actual trading halts, VPIN misses most (above 90%) of the abnormal intraday price movements that ultimately lead to price-limit hits. VPIN limits are therefore imperfect substitutes for price limits. When VPIN-signaled toxic halts are compared with VPIN-signaled non-toxic halts, we find illiquidity and price impact of trades to be more pronounced around the former. This finding is encouraging, as it gives some credibility to VPIN as an early warning signal for truly toxic events.

Unfortunately, when we control for *ex ante* volatility, differences between toxic and non-toxic halts disappear. Yet, there is an exception: VPIN with volume bars. This exception

matters because, first, it is the specification that more closely aligns with ELO's volume-clock paradigm; second, we cannot reject that a well calibrated VPIN can occasionally succeed as an advanced indicator of order flow toxicity. The problem is that it only works intermittently. VPIN with calibrated volume bars has a false positive rate above 80% in signaling order flow toxicity. Also, the most extreme price movements and the worst liquidity conditions that can be found during VPIN-identified toxic periods are similar to those we observe during non-toxic days.

Our study exposes practical difficulties of implementing VPIN limits. As in previous studies (e.g., Andersen and Bondarenko, 2015; Chakrabarty et al., 2015), we find that VPIN is highly sensitive to parameter choices. Our focus is on the type and size of bars, the main BVC parameters. For equities, ELO (2016) recommend to carefully calibrating the BVC stock by stock, but they do not explain how to do it. Following Wu et al. (2013), we calibrate BVC so as to optimize the in-sample forecast power of extreme *RV* realizations. This exercise, however, does not result in a comparable forecast power of extreme illiquidity realizations. Indeed, VPIN with uniform bars works as well as VPIN with calibrated bars. As ELO (2012a) themselves remark, not all volatility is due to toxicity. Hence, a calibration exercise *à la* Wu et al. (2013) might lead to suboptimal results. An interesting and yet unanswered question we leave for future research is whether an in-sample calibrated VPIN performs as well out-of-sample. The answer to this question seems pivotal to gauge the potential of a VPIN-based circuit breaker. What we do show in this paper is that unless it is recalibrated periodically, VPIN could miss toxic events. Determining how often to restart VPIN represents an extra challenge for market authorities.

Our database includes an initiator flag we use to estimate VPIN-flag, that is, VPIN computed using the actual order imbalance (OI). ELO (2012a) claim that VPIN should be independent of the trade-classification algorithm chosen. Our results clearly show that this is

not true. VPIN-flag shows the weakest link with ex-post illiquidity of all the VPIN versions we consider, and is the least efficient in anticipating truly toxic events. Our findings complement Andersen and Bondarenko (2015) that find a negative association between VPIN-flag and posterior short-run volatility. Recently, ELO (2016) advocate for using BVC-based OI estimates. Initiator-based OI, ELO argue, are no longer linked to information-based trading because sophisticated informed traders in modern high-frequency markets use both market and limit orders. This argument is supported by independent studies (e.g., Kim and Stoll, 2014; Collin-Dufresne and Fos, 2015). Even if the aggressor flag becomes distorted in the presence of HFT, this should not be an issue in the first half of our sample. Yet, VPIN-flag does not perform any better during the first half of the sample as compared to the second half. Therefore, we conclude that the ultimate performance of VPIN depends on BVC.

Finally, VPIN may have unintended consequences. VPIN limits themselves could enhance volatility and harm liquidity. For example, in a sort of magnet effect of VPIN limits, as VPIN approaches its limit and a halt starts to look imminent, some traders could advance trades in time, exacerbating imbalances and pushing VPIN towards its limit (e.g., Subrahmanyam, 1994). Liquidity providers that monitor VPIN in real time could also withdraw from the market, exacerbating short-term volatility and the price impact of trades. So far, these problems can only be evaluated theoretically or via experimental markets.

References

Abad, D., Pascual, R., 2010. Switching to a temporary call auction in times of high uncertainty. The Journal of Financial Research 33, 45-75.

Andersen, T.G., Bondarenko, O., 2014. VPIN and the flash crash. Journal of Financial Markets 17, 1-46.

Andersen, T.G., Bondarenko, O., 2015. Assessing measures of order flow toxicity and early warning signals for market turbulence. Review of Finance 19, 1-54.

Angel, J.J., Harris, L.E., Spatt, C.S., 2011. Equity trading in the 21st century. Quarterly Journal of Finance 1, 1-53.

Beber, A., Pagano, M., 2013. Short-selling bans around the world: evidence from the 2007–09 crisis. The Journal of Finance 68, 343-381.

Bhattacharya, A., Chakrabarti, B. B., 2014. An examination of adverse selection risk in Indian IPO after-markets using high frequency data. International Journal of Economic Sciences 3, 1-49.

Borochin, P., Rush, S., 2016. Identifying and Pricing Adverse Selection Risk with VPIN. SSRN Working Paper (http://ssrn.com/abstract=2599871).

Chakrabarty, B., Pascual, R., Shkilko, A., 2015. Evaluating trade classification algorithms: bulk volume classification versus the tick rule and the Lee-Ready algorithm. Journal of Financial Markets 25, 52-79.

Cheung, W. M., Chou, R. K., Lei, A.C.H, 2015. Exchange-Traded Barrier Option and VPIN: Evidence from Hong Kong. Journal of Futures Markets 35, 561–581.

CNMV, 2011. Desarrollos recientes en la microestructura de los mercados secundarios de acciones. Working Paper #50.

Collin-Dufresne, P., Fos, V., 2015. Do prices reveal the presence of informed trading? The Journal of Finance 70, 1555–1582.

Cui, B., Gozluklu, A.E., 2016. Intraday rallies and crashes: spillovers of trading halts. International Journal of Finance & Economics 21, 472-501.

Easley, D., López de Prado, M., O'Hara, M., 2011a. The microstructure of the "Flash Crash": flow toxicity, liquidity crashes, and the probability of informed trading. Journal of Portfolio Management 37, 118–128.

Easley, D., López de Prado, M., O'Hara, M., 2011b. The exchange of flow toxicity. Journal of Trading 6, 8–13.

Easley, D., López de Prado, M., O'Hara, M., 2012a. Flow toxicity and liquidity in a high-frequency world. Review of Financial Studies 25, 1457-1493.

Easley, D., López de Prado, M., O'Hara, M., 2012b. The volume clock: insights into the high frequency paradigm. Journal of Portfolio Management 37, 118-128.

Easley, D., López de Prado, M., O'Hara, M., 2014. VPIN and the flash crash: a rejoinder. Journal of Financial Markets 17, 47-52.

Easley, D., López de Prado, M., O'Hara, M., 2015. Optimal execution horizon. Mathematical Finance 25, 640-672.

Easley, D., López de Prado, M., O'Hara, M., 2016. Discerning information from trade data. Journal of Financial Economics 120, 269-285.

ESMA, 2014. High frequency trading activity in EU equity markets. Economic Report #1.

Foucault, T., Pagano, M., and Röell, A., 2013. Market liquidity: theory, evidence, and policy, Oxford University Press, New York.

Goldstein, M.A., Kavjecz, K.A., 2000. Eighths, sixteenths, and market depth: changes in tick size and liquidity provision on the NYSE. Journal of Financial Economics 56, 125-149.

Hendershott, T., Jones M.C., Menkveld, A., 2011. Does algorithmic trading improve liquidity? The Journal of Finance 66, 1-33.

Holden, C., Jacobsen, S., 2014. Liquidity measurement problems in fast, competitive markets: Expensive and cheap solutions. The Journal of Finance 69, 1747-1785.

Kim, S.T., Stoll, H.R., 2014. Are trading imbalance indicative of private information? Journal of Financial Markets 20, 151-174.

Kyle, A.S., 1988. Trading halts and price limits. Review of Futures Markets 7, 426-434.

Lehmann, B.N., 1989. Commentary: volatility, price resolution, and the effectiveness of price limits. Journal of Financial Services Research 32, 205-209.

Low, R.K.Y., Li, T., Marsh, T., 2016. BV-VPIN: Measuring the impact of order flow toxicity and liquidity on international equities markets. SSRN Working Paper (http://ssrn.com/abstract=2791243).

Madhavan, A., 1992. Trading mechanisms in securities markets. The Journal of Finance 47 607-641.

McInish, T., Upson, J., Wood, R.A., 2014. The Flash Crash: Trading aggressiveness, liquidity supply, and the impact of intermarket sweep orders. Financial Review 49 481-509.

O'Hara, M., 1995. Market microstructure theory, Blackwell, Cambridge, MA.

O'Hara, M., 2015. High frequency market microstructure, Journal of Financial Economics 116, 257-270.

Panayides, M.A., Shohfi, T., Smith, J.D., 2014. Comparing trade flow classification algorithms in the electronic era: The good, the bad, and the uninformative. SSRN Working Paper (http://ssrn.com/abstract=2503628).

Pardo, A., Pascual, R., 2012. On the hidden side of liquidity. European Journal of Finance 18, 949-967.

Pöppe, T., Moss, S., Schiereck, D., 2016. The sensitivity of VPIN to the choice of trade classification algorithm. Journal of Banking & Finance, 73, 165-181.

Phuensane, P., Williams, J.M., 2016. Order flow toxicity and informed trading around known market manipulation events: Evidence from interest rate futures. SSRN Working Paper (http://ssrn.com/abstract=2807531).

Song, J.H., Wu, K., Simon, H.R., 2014. Parameter analysis of the VPIN (volume synchronized of informed trading) metric. In: Zopounidis, C., Galariotis, E. (Eds.), Quantitative Financial Risk Management: Theory and Practice. Wiley.

Subrahmanyam, A., 1994. Circuit breakers and market volatility: a theoretical perspective. The Journal of Finance 49, 237-254.

UK Government Office of Science, 2012. Foresight: The future of computer trading in financial markets: An international Perspective. Final Project Report, London.

Upson, J., Johnson, H., McInish, T.H., 2015. Orders versus trades on the consolidated tape. SSRN Working Paper (http://ssrn.com/abstract=2625401).

Wei, W.C., Gerace, D., Frino, A., 2013. Informed trading, flow toxicity and the impact on intraday trading factors. Australian Accounting, Business and Finance Journal 7, 3-24.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics Bulletin 1, 80-83.

Wu, K., Bethel, W., Gu, M., Leinweber, D., Rübel, O., 2013. A big data approach to analyzing market volatility. Algorithmic Finance 2, 241-267.

Van Ness, B.F., Van Ness, R.A., Yildiz, S., 2016. The role of HFTs in order flow toxicity and stock price variance, and predicting changes in HFTs' liquidity provisions. Journal of Economics and Finance, forthcoming (http://dx.doi.org/10.1007/s12197-016-9374-6).

Zimmermann, K., 2013. Price discovery in European volatility interruptions. SSRN Working Paper (http://ssrn.com/abstract=2365772).

Table 1
Sample statistics.
In Panel A, we provide cross-sectional average daily statistics on trading activity, liquidity, and volatility for our sample of SSE-listed stocks and the subsamples with the ten largest (LC) and ten smallest (SC) stocks by market capitalization at the beginning of each month. *Market capitalization* is in millions of euros. The *relative spread* is the quoted spread prevailing before each trade divided by the quote midpoint. The *quoted depth* is the average of the displayed depth in euros at the best quotes prevailing before each trade. We compute the reported daily averages weighting by trade size. *Trades* is the number of trades registered each day. *€Volume* is the daily volume in euros. We report two proxies of volatility: *High-Low* is the ratio between the highest and the lowest price of the day, and *Realized Volatility* is the daily standard deviation of 1-minute returns. *Price* is the daily average trade price. We provide standard deviations in parenthesis. Panel B provides cross-sectional average changes in trading activity, order flow, and liquidity over the sample period (2002-2013) and for two sub-periods: pre-MiFID (2002-2007) and post-MiFID (2008-2013). *Message traffic* is the sum of all the order submissions, revisions, and cancellations. *Depth(€)* is the average cumulated depth in euros at the five best ask and bid levels. *Dispersion* is the absolute distance between the quote mid-point and the 5th best LOB level. For statistical inference, we use the non-parametric Wilcoxon (1945) test of equality of medians.

Panel A: Cross-sectional average sample statistics

		All (45)	10 largest	10 smallest
Size	Market Cap.	12079.42	24992.82 ***	2392.48
	(x10^{-6})	(15069.08)	(16734.15)	(526.98)
Liquidity	Relative spread	0.1577	0.1241 ***	0.2162
	(x100)	(0.065)	(0.061)	(0.105)
	Depth (€)	4028.01	4696.52 ***	2726.51
	(x10^{-2})	(7972.41)	(6832.76)	(5659.45)
Activity	Trades	1506.55	2286.42 ***	873.10
		(1475.42)	(2050.18)	(480.80)
	€ Volume	3888.18	7687.05 ***	961.80
	(x10^{-4})	(7517.35)	(10634.96)	(462.25)
Volatility	High-Low	2.4991	2.1189 **	2.8748
	(x100)	(0.508)	(0.670)	(1.122)
	Realized Vol	0.7992	0.4631 ***	1.2125
	(x100)	(0.449)	(0.512)	(0.721)
Price		18.00	19.65 *	14.34
		(14.42)	(11.20)	(11.11)

Panel B: Cross-sectional average changes (in %) over the sample period

	2002-2013	Pre-MiFID	Post-MiFID	Dif.
Trades	257.01	46.25	210.76	164.52 ***
Volume (shares)	62.91	38.82	24.09	-14.73 ***
Trade size (shares)	-54.37	-5.08	-49.29	-44.21 ***
Message traffic	1957.95	269.62	1688.33	1418.71 ***
Message traffic/trade	476.45	152.74	323.70	170.96 ***
Relative spread	-56.41	-41.60	-14.81	26.79 ***
Depth (€)	-27.64	273.27	-300.90	-574.17 ***
Dispersion	-65.82	41.14	-106.96	-148.10 ***

***,**,* in Panel A (B) mean that the difference between LC (pre-MiFID) and SC (post-MiFID) is statistically significant at the 1%, 5%, and 10% level, respectively.

Table 2
Toxic events.
We provide cross-sectional median statistics on the number and duration of toxic events from January 2002 to December 2013 for our sample of 45 SSE-listed stocks (interquartile ranges in parenthesis). A toxic event begins when the relative VPIN crosses bottom-up a threshold (p) given by the 99th percentile of the VPIN's empirical CDF; it ends when it crosses up-bottom another threshold (q) that correspond to the 85th percentile. The length of the VPIN's rolling window (n) is 50. We use the actual direction of trades ("flag") and the BVC algorithm to obtain order-imbalance estimates. We report results with time, volume, and trade bars. For each bar type, we provide median statistics across stocks and different bar sizes. We consider time bars from 30 to 1800 seconds in increments of 30 seconds, and trade bars (volume bars) from 1% to 40% of the average daily number of trades (volume bucket size) over the preceding month. The volume bucket size is 1/50th of the average daily volume over the previous month. Persistence in toxicity is measured both in minutes and number of volume buckets. We also report the percentage of toxic events that begin and end within the same trading session. For statistical inference, we use the non-parametric Wilcoxon (1945) rank-sum test of equality of medians.

		Persistence		% intraday
Trade classification	# Events	minutes	buckets	events
BVC - Time bars	30	302	78.3	56.26
	(21.5)	(144.6)	(30.8)	(21.32)
BVC- Volume bars	40	456 ***	55.0 ***	31.25 ***
	(25.5)	(191.6)	(17.3)	(15.76)
BVC- Trade bars	46	429 ***	58.0 ***	38.30 ***
	(27.3)	(124.1)	(14.8)	(14.16)
Flag	7	7297 ***	447.0 ***	0.00 ***
	(9.0)	(9750.6)	(582.5)	(16.67)

***, **, * statistically different from the corresponding time bars' statistic at the 1%, 5%, and 10% level, respectively

Table 3
Robustness of VPIN.
We evaluate the robustness of VPIN when signaling toxic events. In Panel A, we change the bar size within a bar type in applying the BVC approach. In Panel B, we change both the bar type (time, trade, or volume) and the bar size. We randomly pick two different VPIN specifications and compute the proportion of overlapping toxic events. We repeat the process 200 times and obtain the cross-sectional median and interquartile range (in parentheses). Time bars span from 30 to 1800 seconds in increments of 30 seconds. Trade bars vary from 1% to 40% of the of the average daily number of trades over the preceding month (T) in increments of 1%. Volume bars vary from 1% to 40% of the volume bucket size (V) in increments of 1%. The volume bucket size is 1/50 times the average daily volume over the preceding month. In Panel A, we also provide separated statistics for randomly matched VPIN specifications with close and distant bars. We classify two bars as being "close in size" when their size differs less than 5 minutes in case of time bars, or less than 10%T (10%V) in case of trade (volume) bars, respectively. A toxic event begins when the relative VPIN (i.e., the cumulated probability of each VPIN reading) crosses bottom-up the VPIN threshold p=0.99; it ends when it crosses up-bottom a second threshold q=0.85. In computing VPIN, we use a rolling window of 50 volume buckets. For statistical inference, we use the non-parametric rank-sum test of Wilcoxon (1945).

Panel A: Robustness to changes in the bar size (within a bar type)

	All cases	Close in size		Distant in size	
Bar type	% matches		% matches		% matches
Time	0.7072	< 5 min.	0.7508	> 15 min.	0.6509 ‡‡‡
	(0.0674)		(0.0321)		(0.0783)
Volume	0.5835 †††	< 10%V	0.6133	> 20%V	0.5338 ‡‡‡
	(0.0578)		(0.0337)		(0.0449)
Trade	0.6837	< 10%T	0.7731	> 20%T	0.5283 ‡‡‡
	(0.1822)		(0.0597)		(0.0746)

Panel B: Robustness to changes in the bar type

	Time bars	Volume bars	Trade bars
Bar type	% matches	% matches	% matches
Time		0.2910	0.4788 ***
		(0.0480)	(0.0413)
Volume	0.2120		0.2812 ***
	(0.0282)		(0.1184)
Trade	0.3508 ***	0.2850	
	(0.0184)	(0.1488)	

†††, ††, † means different than the time bars case at the 1%, 5%, and 10% level, respectively.

‡‡‡,‡‡,‡ means different than the "close in size" case at the 1%, 5% amd 10% level, respectively.

***, **, * means different than the other statistic in the same row at the 1%, 5% and 10% level, respectively.

Table 4

Illiquidity and volatility around VPIN-limit hits.

We study liquidity and volatility around VPIN-limit hits. A VPIN-limit is hit when the relative VPIN crosses bottom-up the VPIN limit $p = 0.99$. For each VPIN-limit hit, we compare the post-event relative spread (*RS*) and realized volatility (*RV*) with the 95th and 99th percentiles of the empirical distribution of those same metrics over an event-specific benchmark period. Post-event LOB depth (*DK*) is compared with the 5th and 1st percentiles of the same distribution. The benchmark period consists of the 250 days closest to the event day with no toxicity and the same tick regime. We split the benchmark days in regular 5-minute intervals. Liquidity and volatility metrics are standardized using the benchmark mean and standard deviation for the same 5-minute interval. *RS* is the quoted spread divided by the quote midpoint. *DK* is the average between the accumulated euro-volume at the five best bid and offer quotes of the LOB. Liquidity proxies are averaged weighting by time. *RV* is the standard deviation of the 1-minute price changes within each 5-minute interval. We provide the proportion of events with extreme illiquidity and volatility readings for two pre-event and two post-event 5-minute intervals centered on the VPIN-limit violation. We also provide the cross-sectional average deviation of the standardized metric in the pre- or post-event interval from the benchmark percentile. Statistical test are based on the non-parametric Wilcoxon (1945) rank-sum test for equality of medians. In implementing the BVC, we report results for 1-minute time bars, trade bars of 2% of the average daily number of trades over the previous month, and volume bars of 2% of the volume bucket size. The volume bucket size is 1/50th of the average daily volume over the previous month. We also report results for VPIN-flag (VPIN computed using the actual direction of trades).

Panel A: Time bars (60 seconds)

		RS - percentile		*DK* - percentile		*RV* - percentile	
Interval		95th	99th	5th	1st	95th	99th
[-10 -5)	%	18.63	9.80	6.45	3.04	40.36	28.64
	Dif.	*-0.878* ***	*-2.083* ***	*1.534*	*1.715* ***	*0.76* ***	*-0.46* ***
[-5 0)	%	20.30	10.48	5.31	2.65	62.46	49.56
	Dif.	*-0.845* ***	*-2.048* ***	*1.731* ***	*1.910* ***	*4.09* ***	*2.86* ***
[0 5)	%	22.81	12.20	6.17	3.28	51.45	39.10
	Dif.	*-0.735* ***	*-1.940* ***	*1.403*	*1.585* ***	*1.60*	*0.38* ***
[5 10)	%	18.23	9.74	6.12	3.32	41.52	27.80
	Dif.	*-1.041* ***	*-2.248* ***	*1.377* ***	*1.558* ***	*0.57* ***	*-0.65* ***
Panel B: Volume bars (v = 2%)							
[-10 -5)	%	18.31	8.38	8.81	4.38	28.25	17.56
	Dif.	*-0.941* ***	*-2.205* ***	*1.126*	*1.298* ***	*-0.44* ***	*-1.62* ***
[-5 0)	%	21.37	9.30	8.30	4.02	34.95	23.07
	Dif.	*-0.829* ***	*-2.092* ***	*1.235* ***	*1.408* ***	*0.28* ***	*-0.91* ***
[0 5)	%	26.52	12.52	9.01	4.73	42.48	29.57
	Dif.	*-0.643* ***	*-1.910* ***	*1.103* ***	*1.275* ***	*0.72* ***	*-0.47* ***
[5 10)	%	19.62	8.27	9.49	5.58	25.83	15.71
	Dif.	*-1.019* ***	*-2.283* ***	*1.002* ***	*1.175* ***	*-0.74* ***	*-1.93* ***

*** Means the post-event level is different than the benchmark percentile at the 1% level across all toxic events.

Table 4 (Cont.)
Illiquidity and volatility around VPIN-limit hits.

Panel C: Trade bars (v = 2%)

		RS - percentile		*DK* - percentile		*RV* - percentile	
Interval		95th	99th	5th	1st	95th	99th
[-10 -5)	%	24.41	11.96	9.25	4.50	38.20	26.82
	Dif.	*-0.698* ***	*-1.889* ***	*0.976* ***	*1.142* ***	*0.59* ***	*-0.65* ***
[-5 0)	%	23.30	10.97	9.36	4.83	45.93	33.94
	Dif.	*-0.727* ***	*-1.920* ***	*1.059* ***	*1.225* ***	*1.60* ***	*0.37* ***
[0 5)	%	26.04	14.09	9.84	5.01	44.96	31.95
	Dif.	*-0.692* ***	*-1.882* ***	*0.973*	*1.140* ***	*1.01* ***	*-0.22* ***
[5 10)	%	22.99	11.46	9.58	4.85	35.85	23.04
	Dif.	*-0.823* ***	*-2.015* ***	*0.944* ***	*1.112* ***	*0.10* ***	*-1.14* ***
Panel D: Flag							
[-10 -5)	%	6.99	2.62	4.37	1.57	13.64	7.17
	Dif.	*-1.724* ***	*-2.850* ***	*1.981* ***	*2.163* ***	*-1.44* ***	*-2.62* ***
[-5 0)	%	5.27	2.28	3.87	2.28	17.05	9.14
	Dif.	*-1.774* ***	*-2.903* ***	*2.071* ***	*2.252* ***	*-1.31* ***	*-2.48* ***
[0 5)	%	6.63	2.58	4.60	2.39	21.55	12.89
	Dif.	*-1.656* ***	*-2.785* ***	*1.940* ***	*2.122* ***	*-0.76* ***	*-1.94* ***
[5 10)	%	8.72	3.38	4.63	1.60	15.48	7.47
	Dif.	*-1.672* ***	*-2.798* ***	*1.809* ***	*1.991* ***	*-1.37* ***	*-2.55* ***

*** Means the post-event level is different than the benchmark percentile at the 1% level across all toxic events.

Table 5
VPIN-limit vs. price-limit hits.
In Panel A, we provide summary statistics on the number of SSE toxic halts within our sample. A halt is toxic if it falls within a toxic period according to the VPIN metric. A toxic period begins when the relative VPIN (i.e., cumulated probability of each VPIN value) crosses bottom-up the VPIN limit p and ends when it crosses up-bottom a second threshold q. We report results for $p = 0.99$ and $q = 0.85$. In computing VPIN, we use a rolling window of $n = 50$ volume buckets. The volume bucket size equals 1/50 times the average daily volume over the preceding month. In implementing the BVC, we report results for 1-minute time bars, trade bars of 2% of the average daily number of trades over the previous month, and volume bars of 2% of volume bucket size. The volume bucket size is $1/50^{th}$ of the average daily volume over the previous month. We also report results for VPIN-flag, that is, VPIN computed using the actual direction of trades. A halt is static (dynamic) if it is triggered by a violation of the static (dynamic) price limit. Static limits are set over the allocation price of the last auction. Dynamic limits are set over the last trade price. In Panel B, we show the percentage of VPIN-identified toxic periods that comprise at least one trading halt. For those periods, we report the median distance (in minutes) and interquartile range from the VPIN-limit violation to price-limit violation that leads to the closest toxic halt. Finally, we show the proportion of toxic halts that happen before the end of the session where the VPIN-limit violation occurs, and the proportion of toxic halts that happen in the 60-minute interval following the VPIN-limit violation.

Panel A: Toxic trading halts

	Toxic halts		Static halts		Dynamic halts	
VPIN version	Halts	%	halts	%	halts	%
Time bars (60 sec.)	508	7.54	250	8.41	258	6.81
Volume bars (v = 2%)	484	7.19	161	5.47	323	8.52
Trade bars (v = 2%)	654	9.71	344	11.69	310	8.18
Flag	299	4.44	155	5.27	144	3.80

Panel B: Toxic events

	Total	With toxic	Distance to the halt		In the same	Within the next
VPIN version	events	halts (%)	min.	iqr.	session (%)	hour (%)
Time bars (60 sec.)	1364	20.23	78.7	(233.4)	48.82	27.56
Volume bars (v = 2%)	1589	14.60	200.1	(460.7)	24.79	11.98
Trade bars (v = 2%)	1673	19.61	138.5	(408.4)	38.53	20.03
Flag	556	19.1	928.7	(2311.6)	7.69	3.68

Table 6

Toxic halts vs. non-toxic halts.

We test the null hypothesis that market conditions around SSE VPIN-located toxic and non-toxic static halts are alike. Similar results for dynamic halts are reported in the online appendix. Static halts are triggered by violations of the static price limit, set around the allocation price of the last auction. There are 2,943 static halts in our sample. A halt is toxic if it falls within a toxic event according to VPIN. A toxic event begins when the relative VPIN (i.e., the cumulated probability of VPIN values) crosses bottom-up the VPIN limit $p = 0.99$ (VPIN-limit violation) and ends when the relative VPIN crosses up-bottom a second threshold $q = 0.85$. We report results for VPIN with the BVC algorithm with 1-minute time bars (Panel A), volume bars of 2% of volume bucket size (Panel B), and trade bars of 2% of the average daily number of trades over the previous month (Panel C). The volume bucket size is $1/50^{th}$ of the average daily volume over the previous month. The relative spread (*RS*) is the quoted spread divided by the quote midpoint averaged weighting by time. The LOB depth (*DK*) is the accumulated displayed depth at the five best ask and bid quotes in euros also averaged weighting by time. Realized volatility (*RV*) is the standard deviation of the 1-minute price changes. The price impact (*PI*) is the difference between the quote midpoint one minute after the trade and the quote midpoint prevailing before the trade, averaged weighting by trade size. All the metrics are standardized by subtracting the mean and dividing by the standard deviation of the corresponding variable for the exact same time interval across 250 non-toxic benchmark days with no trading halts. We use Wilcoxon's (1945) rank-sum test to provide statistical significance to differences between toxic and non-toxic halts. We use a 30-minute window centered on the trading halt that we split into five-minute intervals.

Panel A: Static halts and VPIN with time bars (60 sec.)

	RV	*RS*	*DK*	*PI*
[-15 -10)	2.019 ***	0.766 ***	0.107	1.074 ***
[-10 -5)	2.840 ***	0.965 ***	0.072	1.720 ***
[-5 0)	2.766 ***	0.942 ***	0.092	2.084 ***
[0 5)	4.956 ***	1.197 ***	0.188	1.769 ***
[5 10)	2.437 ***	1.539 ***	0.059	1.062 ***
[10 15)	1.536 ***	1.022 ***	0.021	1.075 ***

Panel B: Static halts and VPIN with volume bars (v = 2%)

[-15 -10)	1.285 **	1.676 ***	-0.258 ***	1.579 ***
[-10 -5)	1.347 ***	1.736 ***	-0.240 ***	2.610 ***
[-5 0)	1.878 ***	1.587 ***	-0.333 ***	4.117 ***
[0 5)	4.612 ***	1.792 ***	-0.153 ***	2.978 ***
[5 10)	1.771 ***	1.583 ***	-0.184 ***	1.782 ***
[10 15)	0.516	1.448 ***	-0.202 ***	0.951

Panel C: Static halts and VPIN with trade bars (v = 2%)

[-15 -10)	1.466 ***	0.732 ***	-0.154 ***	1.229 ***
[-10 -5)	1.758 ***	1.027 ***	-0.198 ***	1.360 ***
[-5 0)	2.638 ***	1.138 ***	-0.192 ***	2.278 ***
[0 5)	3.733 ***	1.272 ***	-0.034 **	2.034 ***
[5 10)	1.839 ***	1.474 ***	-0.128 ***	1.468 ***
[10 15)	1.630 ***	1.193 ***	-0.143 ***	0.953 ***

Panel D: Static halts and VPIN-flag

[-15 -10)	0.312 **	-0.133	0.247 **	-0.121
[-10 -5)	0.202	-0.184	0.188 ***	-0.017
[-5 0)	0.419 *	-0.126	0.221 ***	-0.411
[0 5)	0.181	-0.060	0.284 **	0.029
[5 10)	-0.010	-0.067	0.224 **	-0.290
[10 15)	0.236	-0.084	0.215	-0.480

***, **, * means statistically significant at the 1%, 5%, and 10%, respectively.

Table 7

Toxic halts vs. non-toxic halts: regression analysis.

We test the null that differences in market conditions between toxic and non-toxic halts are driven by differences in volatility. A halt is toxic when it falls within a toxic event, which begins when the relative VPIN (i.e., the cumulated probability of VPIN values) crosses bottom-up the VPIN limit $p = 0.99$ and ends when it crosses up-bottom a second threshold $q = 0.85$. We use data on 6,734 trading halts for 45 SSE-listed stocks from 2002 to 2013. We distinguish between static halts (Panel A) and dynamic halts (Panel B). Static (dynamic) halts are triggered by violations of price limits set around the allocation price of the last auction (the last trade price). There are 2,943 static halts and 3,791 dynamic halts. We provide results for VPIN-flag, that is, VPIN computed with the actual trade direction. We also provide results for VPIN with the BVC algorithm and three types of bars: 60-second time bars, 2% volume bars, and 2% trade bars. Dependent variables are the average relative spread (*RS*), the average LOB depth (*DK*), both weighted by time, and the average price impact of trades (*PI*) weighted by trade size. Explanatory variables are the first three lags of realized volatility (*RV*) – the standard deviation of the 1-minute price changes, a dummy variable for toxic halts (*Toxic*), and the stock-specific price range at the time of the halt (*Range*), either static or dynamic. As controls, we include 11 yearly dummies and 44 stock dummies. We only report the estimated coefficient of the variable of interest: *Toxic*. The model is defined for five-minute intervals around trading halts. We report estimated coefficients for the first interval preceding and the first interval following the halt. The model is estimated by OLS with White-robust standard errors.

Panel A: Static halts

	RS		*DK*		*PI*	
Toxic coef.	[-5,0)	(0,5]	[-5,0)	(0,5]	[-5,0)	(0,5]
Time bars	0.140	0.382	0.242 ***	0.255 **	1.334	0.823
Adj.-R^2	0.229	0.123	0.094	0.082	0.106	0.071
F-test	17.200	15.146	10.560	7.231	12.843	7.836
Volume bars	0.702 ***	0.687 ***	-0.003	0.111	2.580 *	1.913 **
Adj.-R^2	0.241	0.124	0.111	0.096	0.105	0.078
F-test	17.983	16.290	16.023	10.451	14.203	8.068
Trade bars	0.553 ***	0.566 **	-0.036	0.108	1.932 *	1.239 **
Adj.-R^2	0.234	0.139	0.109	0.097	0.104	0.069
F-test	17.561	15.284	11.213	10.108	13.563	7.543
Flag	-0.128	-0.173	0.143	0.234 *	-0.753	-0.110
Adj.-R^2	0.233	0.131	0.094	0.080	0.110	0.067
F-test	17.637	14.498	10.748	6.655	13.880	7.237
Obs.	2821	2835	2821	2835	2760	2787
Panel B: Dynamic halts						
Time bars	0.769 *	1.221 ***	0.151 **	0.298 ***	1.132	1.850 **
Adj.-R^2	0.238	0.203	0.007	0.060	0.027	0.182
F-test	13.867	17.138	2.430	7.240	2.977	5.920
Volume bars	0.921 **	1.107 ***	-0.034	0.133	2.668 **	2.022 **
Adj.-R^2	0.254	0.217	0.008	0.068	0.030	0.193
F-test	16.861	21.914	6.319	11.440	3.039	5.808
Trade bars	0.166	0.831 *	-0.032	0.092	2.066 *	0.736
Adj.-R^2	0.243	0.214	0.008	0.054	0.028	0.207
F-test	12.278	15.712	2.091	6.679	3.624	6.024
Flag	0.172	0.471	-0.003	0.143	3.749	-0.234
Adj.-R^2	0.197	0.164	0.006	0.055	0.039	0.142
F-test	13.276	16.366	2.022	6.391	3.413	5.276
Obs.	2844	2953	2844	2953	2557	2619

***, **,* means statistically significant at the 1%, 5%, and 10% level, respectively.

Table 8
Restarting VPIN.
We study the sensitivity of VPIN to the starting point of the series. We consider two starting points: January 2002 ("VPIN02") and January 2009 ("VPIN09"). For each case, we obtain the number of toxic events and toxic halts between 2009 and 2013. A toxic event begins when the relative VPIN (i.e., the cumulated probability of each VPIN observation) crosses bottom-up the VPIN limit $p = 0.99$ (VPIN-limit violation) and ends when it crosses up-bottom another threshold $q = 0.85$. For both toxic events and toxic halts, we provide the total number of events for the whole sample ("All") and the average number of events across stocks ("Avg."). We use the non-parametric Wilcoxon's rank-sum test to compare VPIN02 and VPIN09. Finally, we provide the proportion of VPIN02-identified toxic halts also pinpointed by VPIN09, and the proportion of VPIN09-identified toxic halts located by VPIN02 too. We apply the BVC algorithm to assign direction to trades. We report results implementing BVC using 60-second time bars, 2% volume bars (over the volume bucket size), and 2% trade bars (over the average number of trades of the asset over the previous month). The volume bucket size is $1/50^{th}$ of the average daily volume over the previous month. We also report results using VPIN-flag, that is, VPIN computed using the actual direction of trades.

	Toxic events				Toxic halts				Matched toxic halts (%)	
	VPIN02		VPIN09		VPIN02		VPIN09			
VPIN version	All	Avg.	All	Avg.	All	Avg.	All	Avg.	VPIN02	VPIN09
Time bars (60 sec.)	349	7.76	591	13.13 ***	56	1.24	83	1.84 **	89.29	60.24
Volume bars (v = 2%)	142	3.16	733	16.29 ***	18	0.40	54	1.20 ***	100.00	33.33
Trade bars (v = 2%)	318	7.07	831	18.47 ***	81	1.80	150	3.33 ***	95.06	51.33
Flag	157	3.49	217	4.82	74	1.64	65	1.44	59.46	67.69

***, **, * means that the average per stock with VPIN09 is different than with VPIN02 at the 1%, 5% and 10% level, respectively.

Table 9

Static and dynamic price variations within toxic periods.

This table provides summary statistics (cross-sectional mean and standard deviation) on the maximum dynamic (Panel A) and static (Panel B) price variations within toxic periods according to the VPIN. We also provide the distribution of dynamic (Panel A) and static (Panel B) price variations with respect to SSE pre-established categories of dynamic and static ranges. A dynamic price variation is the relative change in prices with respect to the previous transaction price. A static price variation is the relative change in prices with respect to the static price, that is, the allocation price of the last auction completed. We also provide the average 99th percentile and the maximum of both dynamic and static price variations across benchmark periods. Each benchmark period is event-specific, and given by the 250 days closest to the toxic event with no remarkable toxicity, no trading halts, and the same tick regime. A toxic event begins when the relative VPIN (i.e., the cumulated probability of each VPIN observation) crosses bottom-up the VPIN limit $p = 0.99$ (VPIN-limit violation) and ends when it crosses up-bottom another threshold $q = 0.85$. We consider different VPIN implementations. We report results using the BVC algorithm with 60-second time bars, 2% volume bars (over the volume bucket size), and 2% trade bars (over the average number of trades of the asset over the previous month). The volume bucket size is 1/50th of the average daily volume over the previous month. We also report results using VPIN-flag, that is, VPIN computed using the actual direction of trades.

Panel A: Price changes with respect to the dynamic price

	Toxic periods		Benchmark period		Distribution of max. dynamic price change (%) w.r.t. stadardized SSE dynamic limits			
VPIN version	Max.	Std.	99th	Max.	<=1% (min)	<=2%	<=3%	<=4%
Time bars (60 sec.)	0.0072	0.0051	0.0040 ***	0.0191 ***	79.67	97.60	99.63	99.91
Volume bars (v = 2%)	0.0072	0.0048	0.0042 ***	0.0176 ***	78.82	98.01	99.63	99.93
Trade bars (v = 2%)	0.0082	0.0058	0.0040 ***	0.0203 ***	73.97	95.51	99.03	99.93
Flag	0.0083	0.0047	0.0049 ***	0.0211 ***	70.76	98.66	99.55	100.00

Panel B: Price changes with respect to the static price

	Toxic periods		Benchmark period		Distribution of max. static price change (%) w.r.t. standardized SSE static limits			
VPIN version	Max.	Std.	99th	Max.	<=4% (min)	<=5%	<=7%	<=8%
Time bars (60 sec.)	0.0282	0.0175	0.0364 ***	0.0567 ***	81.33	91.59	97.50	98.61
Volume bars (v = 2%)	0.0224	0.0137	0.0395 ***	0.0604 ***	91.88	97.34	99.19	99.56
Trade bars (v = 2%)	0.0290	0.0182	0.0387 ***	0.0613 ***	79.58	89.75	96.34	97.83
Flag	0.0277	0.0199	0.0426 ***	0.0674 ***	82.14	89.06	95.76	97.54

*** means statistically different from the cross-sectional average maximum price variation within toxic events a the 1% level.

Table 10
Highest relative spread realizations within toxic periods.
This table provides cross-sectional average statistics about extreme relative spread realizations during both VPIN-identified toxic periods and non-toxic (benchmark) days. A toxic event begins when the relative VPIN (i.e., the cumulated probability of each VPIN observation) crosses bottom-up the VPIN limit $p = 0.99$ (VPIN-limit violation) and ends when it crosses up-bottom another threshold $q = 0.85$. We split each toxic period into five-minute intervals, and compute the standardized relative spread weighted by time (*RS*) within each interval. We pick the maximum *RS* across intervals, which we compare with the same statistic, for the same interval, but over the corresponding benchmark days. The event-specific benchmark period is given by the 250 days closest to the toxic event with no remarkable toxicity, no trading halts, and the same tick regime. Each *RS* observation is standardized using the *RS* mean and standard deviation across the benchmark days for the same five-minute interval. We report standardized metrics in Panel A, and the corresponding raw metrics in basis points in Panel B. We also provide the percentage of toxic events for which the maximum *RS* is below the benchmark readings. We compute VPIN with BVC-based order imbalance estimates. We report results using the BVC algorithm with 60-second time bars, 2% volume bars (over the volume bucket size), and 2% trade bars (over the average number of trades of the asset over the previous month). The volume bucket size is $1/50^{th}$ of the average daily volume over the previous month. We also report results using VPIN-flag, that is, VPIN computed using the actual direction of trades.

Panel A: Standardized *RS* (in standard deviations from the benchmark mean)

	Toxic periods		Benchmark periods	
VPIN version	Max.	Std.	Max.	< bench.(%)
Time bars (60 sec.)	2.60	3.22	5.07 ***	87.89
Volume bars (2%)	3.30	3.32	5.02 ***	81.80
Trade bars (2%)	3.64	4.17	4.89 ***	77.29
Flag	3.18	4.58	4.77 ***	81.98

Panel B: Raw *RS* values (in basis points)

	Toxic periods		Benchmark periods	
VPIN version	Max.	Std.	Max.	< bench.(%)
Time bars (60 sec.)	42.76	37.46	69.74 ***	84.64
Volume bars (2%)	51.07	41.06	69.55 ***	79.04
Trade bars (2%)	49.33	41.36	65.23 ***	77.74
Flag	51.46	41.21	75.23 ***	82.65

***, **,* Means statistically different from the toxic periods a the 1%, 5% and 10% level, respectively.

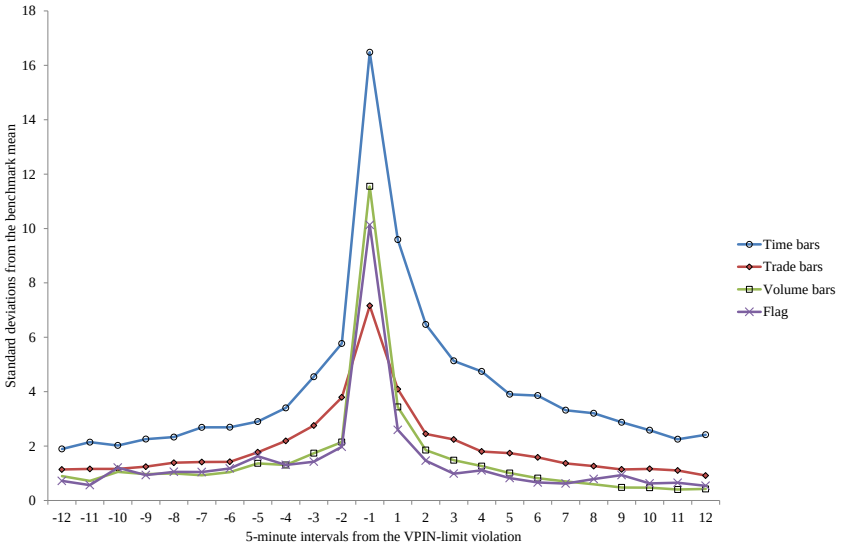
Table 11
Toxic halts vs. non-toxic halts with calibrated bars.
We use a regression approach to test the null that differences in market conditions between toxic and non-toxic halts are driven by differences in volatility. A halt is toxic when it falls within a toxic event, which begins when the relative VPIN (i.e., the cumulated probability of VPIN values) crosses bottom-up the VPIN limit $p = 0.99$ and ends when it crosses up-bottom a second threshold $q = 0.85$. We use data on 6,734 trading halts for 45 SSE-listed stocks from 2002 to 2013. We distinguish between static halts (Panel A) and dynamic halts (Panel B). Static (dynamic) halts are triggered by violations of price limits set around the allocation price of the last auction (the last trade price). There are 2,943 static halts and 3,791 dynamic halts. We use calibrated time, volume, and trade bars. Dependent variables are the average relative spread (*RS*), the average LOB depth (*DK*), both weighted by time, and the average price impact of trades (*PI*) weighted by trade size. As explanatory variables we use the first three lags of realized volatility (*RV*) – the standard deviation of the 1-minute price changes, a dummy variable for toxic halts (*Toxic*), and the stock-specific price range at the time of the halt (*Range*), either static or dynamic. As controls, we include 11 yearly dummies and 44 stock dummies. We only report the estimated coefficient of the variable of interest: *Toxic*. The model is defined for five-minute intervals around trading halts. We report estimated coefficients for the first interval preceding and the first interval following the halt. The model is estimated by OLS with White-robust standard errors.

Panel A: Static halts

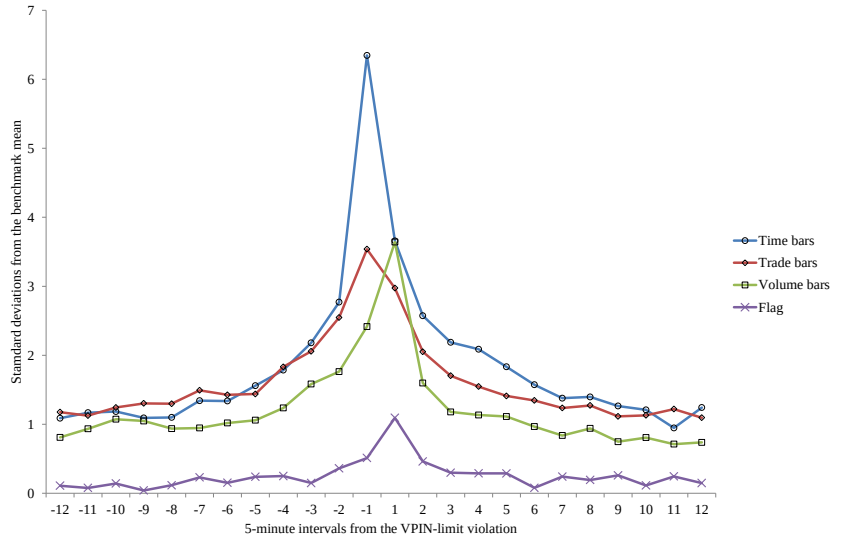
	RS		*DK*		*PI*	
Toxic coef.	[-5,0)	(0,5]	[-5,0)	(0,5]	[-5,0)	(0,5]
Time bars	-0.126	-0.234	0.172 **	0.192 **	-0.216	0.232
Adj.-R^2	0.217	0.115	0.089	0.080	0.101	0.068
F-test	17.426	14.920	9.446	6.918	13.780	7.509
Volume bars	0.831 ***	0.385 *	-0.169 ***	-0.025	1.633	1.375 **
Adj.-R^2	0.247	0.140	0.122	0.106	0.105	0.069
F-test	19.128	17.052	15.367	10.959	13.776	8.301
Trade bars	0.076	-0.044	0.044	0.019	0.331	0.148
Adj.-R^2	0.219	0.133	0.104	0.098	0.105	0.064
F-test	17.745	13.482	10.623	9.454	13.478	7.148
Obs.	2821	2835	2821	2835	2760	2787
Panel B: Dynamic halts						
Time bars	0.444	0.944	0.190 **	0.243 **	0.320	-0.064
Adj.-R^2	0.236	0.203	0.006	0.051	0.027	0.169
F-test	12.358	14.864	1.914	5.965	3.267	6.195
Volume bars	1.218 ***	1.372 ***	-0.134	0.012	2.484 **	1.983 ***
Adj.-R^2	0.268	0.224	0.009	0.064	0.029	0.191
F-test	16.276	19.816	6.721	11.545	3.334	6.261
Trade bars	0.194	0.848 *	0.030	0.123	-0.052	-0.062
Adj.-R^2	0.234	0.205	0.008	0.056	0.026	0.149
F-test	12.686	15.870	2.278	6.320	3.022	5.159
Obs.	2844	2953	2844	2953	2557	2619

***, **,* means statistically significant at the 1%, 5%, and 10% level, respectively.

Figure 1
Abnormal liquidity, volatility, and activity patterns around VPIN-limit hits.
We plot average abnormal liquidity, volatility and trading activity levels around VPIN-limit violations for our sample of 45 Spanish SIBE-listed stocks from January 2002 to December 2013. A toxic event begins when the relative VPIN (i.e., the cumulated probability of the VPIN) crosses bottom-up the $p = 0.99$ limit ("VPIN-limit violation"). Order imbalances are estimated using the BVC algorithm with 60-second time bars, 2% volume bars (over the volume bucket size) and 2% trade bars (over the average number of trades of the asset over the previous month). The volume bucket size is $1/50^{th}$ of the average daily volume over the previous month. For each event, we take the closest 250 non-toxic days as the benchmark period. We consider twenty-four 5-minute intervals centered on the VPIN-limit violation. Each 5-minute interval observation is standardized by subtracting the mean and dividing by the standard deviation of the same metric during the exact same interval across all the benchmark days. We plot averages across all VPIN-limit violations. Statistically significant abnormal values are highlighted using different markers per variable. Statistical tests are based on the Wilcoxon (1945) signed rank-sum test. We use volume in shares (*VOL*) to proxy for trading activity – Figure 1.a. We measure realized volatility as the standard deviation of the 1-minute changes in trade prices (*RV*) – Figure 1.b. Finally, we use the relative spread (*RS*) – Figure 1.c, and the displayed LOB depth in euros at the best quotes (*DB*) – Figure 1.d, all weighted by time, as inverse and direct proxies for liquidity, respectively.

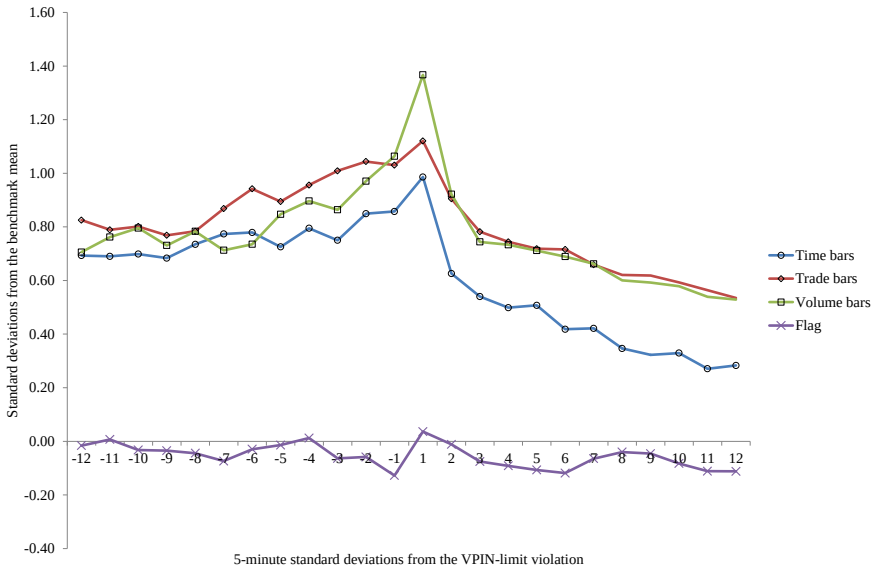


1.1. Volume (shares)

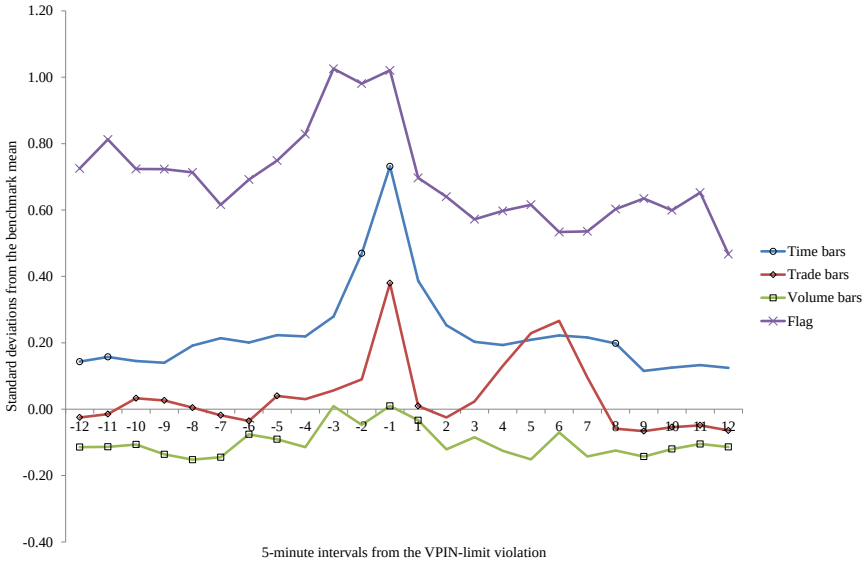


1.2. Realized volatility

Figure 1 (Cont.)
Abnormal liquidity, volatility, and activity patterns around VPIN-limit hits.

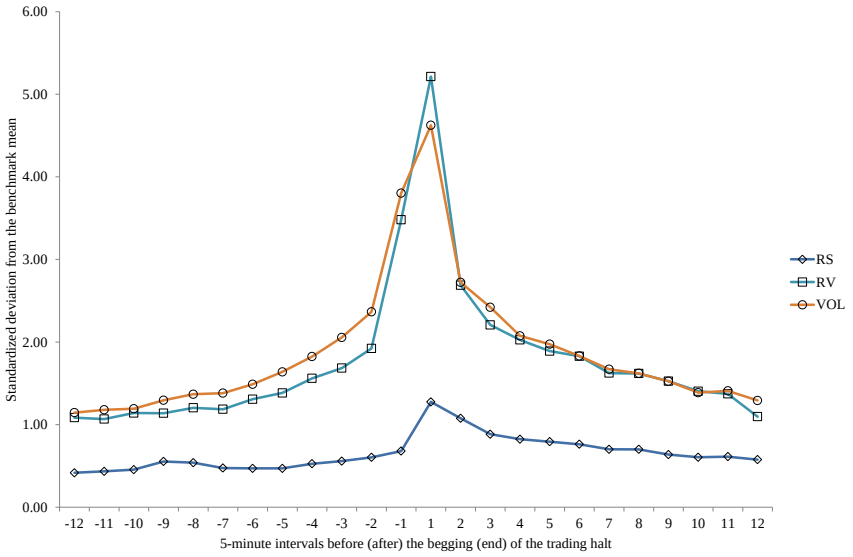


1.3. Relative bid-ask spread

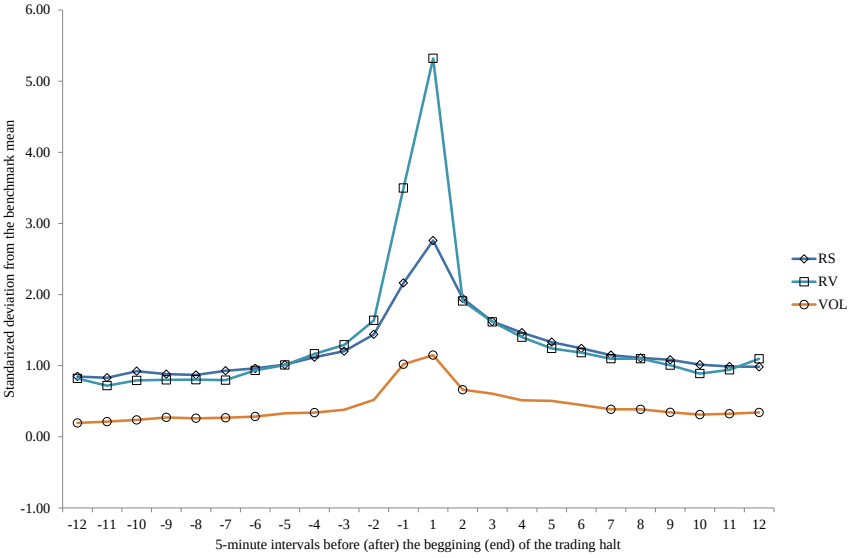


1.4. Euro-depth at the market quotes

Figure 2
Abnormal liquidity, volatility, and activity patterns around price-limit hits.
We plot the average abnormal relative spread (*RS*), realized volatility (*RV*) and volume (*VOL*) around SSE single-stock trading halts, known as "volatility auctions". Our sample consists of 45 Spanish SIBE-listed stocks from January 2002 to December 2013. We distinguish between static (Figure 2.a) and dynamic (Figure 2.b) halts. A static halt is triggered by a violation of the static price limits, the maximum variation permitted around the static price. The static price is the allocation price of the last auction completed. A dynamic halt is triggered by a violation of the dynamic price limits, the maximum variation allowed around the last trade price. We consider twelve five-minute intervals before the price limit hit and twelve five-minute intervals after the resumption of the continuous session. *RS* is the bid-ask spread divided by the quote midpoint; *RV* is the standard deviation of the one-minute changes in prices, and *VOL* is measured in shares. For each event, we take the 250 days closest in time with the same tick regime and no trading halts as the benchmark period. Each observation is standardized by subtracting the mean and dividing by the standard deviation of the same variable for the exact same time interval across the 250 benchmark days. We plot averages across all trading halts. Statistically significant abnormal values are highlighted using different markers per variable. Statistical significance is evaluated using Wilcoxon (1945) non-parametric test.



2.1. Static halts



2.2. Dynamic halts

Appendix:
Parameterization of VPIN and BVC

We summarize our parameter choices for implementing the order flow toxicity measure VPIN with the bulk-volume classification (BVC) algorithm.

VPIN	Description	Alternatives considered
n	Number of volume buckets in each update of VPIN	25, 50, 75
δ	Percentage of the daily average volume	1/50
V_i	Size of the volume bucket for stock *i*	δ times the average daily trading volume of the asset over the preceding month, rounded to the closest integer.
$OI_{i\tau}$	Order imbalance for stock *i* in the τ-th volume bucket	BVC-based order flow imbalance OI based on the actual direction of trades
BVC		
Bar types	Data is pre-aggregated into bars of equal size	Time, trade and volume bars.
Bar size	Pre-determined bar sizes	Time bars (in seconds): From 30 to 1800, in increments of 30 seconds (i.e., 60 different bars). Trade bars (in number of trades): Stock-specific: from 1% to 40%, in increments of 1%, of the average daily number of trades over the preceding month, rounded to the closest integer (i.e., 40 different bars) Volume bars (in shares traded): Stock-specific: from 1% to 40%, in increments of 1%, of 1/50th of the average daily volume (in shares) over the preceding month, rounded to the closest integer (i.e., 40 different bars)
$\Phi(\cdot)$	CDF of the probabilistic distribution assumed for the standardized changes in prices between two consecutive bars	t-student with 0.1 and 0.25 degrees of freedom Normal
$[p, q]$	Thresholds for the CDF of VPIN	[0.99,0.85], [0.95,0.85]