Machine Learning from Randomized Experiments: The Case of the Tick Size Pilot Program*

Hyungil Kye†
Aalto University School of Business

October 2020

Abstract

This paper investigates the Tick Size Pilot Program with the goal of policy evaluation beyond average treatment effect. Using a machine learning approach, I study policy effects stock-by-stock on three major market quality measures, percentage quoted spread, consolidated displayed depth, and high-low volatility. For each pilot stock, I test whether it receives significant treatment effects. I find less than half of the pilot stocks in the treatment groups show positive significance for percentage quoted spread; more than 80% shows positive significance for consolidated displayed depth; only less than 5% shows significance for high-low volatility in either direction; the control group stocks rarely show significance for all the outcomes, revealing no spillover effect at the individual level. Tick constrainedness turns out to be useful in explaining differing significance only for percentage quoted spread, but not for consolidated displayed depth. Percentage realized spread, though, appears to explain for the both outcomes: the lower percentage realized spread, the more likely is the null hypothesis rejected, indicating less-profitable stocks for liquidity providers in the pre-intervention periods tend to receive significant effects in the post-intervention periods.

JEL Codes: C93, G12, G14, G18

Keywords: Tick Size Pilot Program, Applied Machine Learning, Randomized Controlled Trial.

*I am indebted to Bruce Mizrach for his thoughtful guidance throughout my PhD study. I thank Yenjae Chang and Roger Klein for their helpful comments. I gratefully acknowledge the Office of Advanced Research Computing at Rutgers University for providing access to the Amarel cluster and associated research computing resources that have contributed to the results reported here. Bruce Mizrach also generously supported data storage for this project. All errors remain my own responsibility.

†Hyungil "Henry" Kye is a postdoctoral researcher in the Department of Finance at Aalto University School of Business. Email: henry.kye@aalto.fi.

1. Introduction

Randomized experiments are becoming a crucial instrument of empirical studies in economics and finance. Over the past decade, randomized controlled trials (RCTs) have played a special role in policy evaluations in a wide scope of subjects. Figure 1, drawn from Currie, Kleven, and Zwiers (2020), shows a rising popularity of the use of RCT in the area of applied microeconomics. Since 2005, research papers written in the RCT framework among the National Bureau of Economic Research (NBER) working papers have been growing significantly, and a similar trend is also seen in the top-five academic journals of economics.[1]

[Figure 1 about here]

The basic idea of RCT strategy for policy evaluation is the following. Taking policy intervention of interest as *treatment*, a typical RCT carries individuals, firms or communities assigned to treatment and control groups at random.[2] In a well-designed case, average difference of outcome of interest between treatment and control groups defines an *unbiased* estimator of average treatment effect (ATE), which plays a central role in policy learning via RCT.[3] As in cases of clinical trials, such RCT approaches are widely believed to be the gold standard for effectiveness research, delivering "more credible" results on policy problems.

Yet, there is voice of caution, a counterweight to the enthusiasm for the use of RCT to policy problems (e.g., Deaton, 2019; Deaton and Cartwright, 2018; Heckman, 2008, 2020). Among many concerns raised in economics is that ATE, for which RCTs are basically designed, might not be as useful as many believe so. This point is recently highlighted in the RCT literature: "At best, an RCT yields an unbiased estimate, but this property is of limited practical value" (Deaton and Cartwright, 2018, Abstract), and "Advocates of randomization implicitly assume that certain mean differences in outcomes are invariably the objects of interest in performing an [policy] evaluation" (Heckman, 2020, p. 9).

What makes RCTs special is due to the statistical property that researchers can estimate ATE "cleanly" under minimal assumptions. As far as economics is concerned, however, knowing ATE neatly does not necessarily lead to good policy learning. At best, ATE can deliver policy implications around average.[4] At worst, though, ATE could mislead overall policy effects. It is not rare that individuals exposed to policy intervention under a RCT reveal heterogeneous responses. In such a case, it is likely that ATE falls short of representing individual (unit-level) treatment effects (ITEs) meaningfully regardless of how precise it is.[5] Deaton and Cartwright (2018) further point out that statistical inference on ATE becomes less reliable with asymmetric distribution of ITEs.[6] Recently, Young (2019) investigates 53 randomized

[1] *American Economic Review, Econometrica, Journal of Political Economy, Quarterly Journal of Economics*, and *Review of Economic Studies*.

[2] For a formal description of RCT in economic contexts, see Athey and Imbens (2017).

[3] The practice of RCT has been particularly popular in development economics. For summary of RCT-based studies in this area, see Banerjee, Duflo, and Kremer (2016). In market microstructure, Boehmer, Jones, and Zhang (2020), among others, stress at the beginning of the paper, "To gauge the effects of a new regime, a particularly useful approach for the regulator is to test out a new policy by conducting a randomized experiment."

[4] For instance, "The case for randomization is weaker if the analyst is interested in other measures of the distribution or the distribution itself. In general, randomization is not an effective procedure for identifying median gains, or the distribution of gains or many other key parameters" ((Heckman, 2008, p. 12)). Similarly, "RCTs are informative about the mean of the treatment effects, but do not identify other features of the distribution" ((Deaton, 2010, p. 439)).

[5] For example, "The trial might reveal an average positive effect although nearly all of the population is hurt with a few receiving very large benefits, a situation that cannot be revealed by the RCT" ((Deaton, 2010, p. 463)).

[6] "Even less recognized are problems with statistical inference, and especially the threat to significance testing posed when there is an asymmetric distribution of individual treatment effects in the study population" ((Deaton and Cartwright, 2018, p. 2)).

experiment papers published in academic journals of the American Economic Association, and find non-trivial cases of overstated significance of ATE, or its variants, in comparison with testing results drawn from randomization inference.

[Figure 2 about here]

Meanwhile, the use of big data and machine learning (ML) in empirical economic studies has been surging. Figure 2 shows a skyrocketing trend in recent years among NBER and top-five economic journal papers adopting them in the area of applied microeconomics. A considerable part of ML applications to economic studies in particular are focused on causal inference.[7] On this front, ML techniques that offer outstanding out-of-sample predictions are employed at an intermediate stage of casual analysis. In short, those are deployed in predicting counterfactuals with high accuracy, helping researchers overcome the *fundamental problem of causal inference*.[8]

A major difference of such ML approaches with a conventional way of causal inference is that there is no necessarily involvement of a control group but a data-driven predictive model imputing counterfactual, which instead of control group offers the baseline outcome for comparison. Hal R. Varian assesses the potential of ML-based casual inference as, "A good predictive model can be better than a randomly chosen control group, which is usually thought to be the gold standard" ((Varian, 2014, p. 24)).[9] Importantly, causal analysis in this ML way is seamlessly applicable to RCT so long as there are big data backing up the ML plan. One potential gain of introducing ML to RCT is that researchers can tackle directly ITE, learning policy effects from ITE rather than ATE, a summary of it.

The Tick Size Pilot Program (TSPP) is the latest RCT conducted by regulatory bodies in the U.S. stock market over Oct. 2016–Sept. 2018. It is intended to understand the impact of an increase in tick size, the minimum increment in price grid of quotation, from \$0.01 to \$0.05 on market outcomes for small-cap (\$3 billion or less) stocks. While there have been many of the events of tick-size changes in history of the U.S. stock market structure, such as the tick-size decrease in 1997 from \$1/8 to \$1/16 and another decrease in 2001 from \$1/16 to \$0.01, TSPP is unique in that it addresses an event of a tick size *increase* in the form of RCT.

[Table 1 about here]

The response of academia to TSPP has been quite huge. At the time of writing this paper, I find at least 17 academic papers, available as online working papers or published in well-known academic journals, that all exploit TSPP for dealing with their research questions, as listed in Table 1. While different papers carry different research questions as observed by diverse choice of outcomes, the table also shows that the empirical approaches of nearly all the papers converge to Difference-in-Difference (DiD).[10] Basically DiD delivers ATE in panel-data, just as the pooled mean differences of outcomes between treatment and

[7] For recent discussions of ML applications in a general context of economics, see Athey (2017), Athey and Imbens (2019), Kleinberg et al. (2015), Mullainathan and Spiess (2017), and Varian (2014).

[8] The fundamental problem of causal inference refers to the statistical problem that in the potential outcome framework of causal inference, researchers cannot observe counterfactuals, for example, the hypothetical outcomes of treated units in absence of treatment during treatment period. For detail, see Holland (1986).

[9] This basic idea of ML extension to causal analysis is philosophically is similar to that of the *Synthetic Control Method* (SCM) that specializes in comparative case study at an aggregate level. In SCM, a synthetic control, an artificial control unit constructed in data-driven ways, provides a baseline outcome for comparison with actual, intervened outcome for the treated unit. For introduction, see Abadie (2019, forthcoming).

[10] Because there is no meaningful reason for differentiation, I count as DiD the two-way fixed-effect panel-data models with treatment binary indicators on groups and periods.

control groups.[11] In light of the earlier discussions led by, among others, Angus Deaton and James J. Heckman, then, it may be the case that the prior works on TSPP have depicted only limited aspects of policy effects and that some of existing conclusions could suffer from false positive, as noted by Young (2019).

Many of the prior works on TSPP perform a subgroup analysis, mostly in a similar manner of Difference-in-Difference-in-Difference (DDD), in attempt to capture policy effect heterogeneity. However, there has been *only one* characteristic put in place among those collective efforts, namely, tick constrainedness.[12] Moreover, while such analyses are believed to be "concise" as is ATE because they are implemented in a RCT, post-trial subgroup analyses in interacted regression models, like DDD, do not automatically benefit from statistical advantage of RCT, unlike to ATE.[13] As detailed in Deaton (2010, Section 4.2), such practices accompany more assumptions for unbiasedness than the trial itself.

On this background, I investigate TSPP with the main goal of taking policy evaluations beyond ATE. For the sake of brevity, I focus only on three outcomes that are among the most representative market quality measures in the related literature: percentage quoted spread, consolidated displayed depth, and high-low volatility that approximates a short-term volatility. For each outcome I perform hypothesis testing stock-by-stock to see whether a pilot stock i in TSPP reveals significant policy effects on it due to implementation of TSPP. This in analogy is similar to a single parameter testing carried out one-by-one in a multiple-variable linear regression model. In the TSPP context, this approach is particularly useful because it can address at the individual level the issue of spillover effects raised in several prior works that the pilot stocks in the control group also experience a certain degree of the treatment effects. Lastly, I look into policy effect heterogeneity based on testing results on ITEs. Knowing significance of policy effects at the individual level enables to conduct heterogeneity analysis in a more extensive manner, unlimited to a single covariate.

To this end, I propose a novel ML procedure exploiting publicly available large-scale quote data for the U.S. stock market. In the proposed ML procedure, quote data in the whole year of 2015 are summarized over every half-hour segment during the regular trading session, 09:35 - 15:55, excluding the first and last five minutes. Those big data then are put in place to train a ML model for each outcome stock-by-stock in a similar manner of *Synthetic Control Method* (SCM); that is, the outcomes for thousands of U.S.-traded stocks *outside* TSPP constitute the right-hand-side of the ML model along with other variables exogenous of TSPP, such as VIX, time fixed effects at multiple levels, and their interactions. Having being trained via Elastic Net regression, it predicts the outcome for the nine months before and after Oct. 2016, the policy phase-in month, and derives the prediction errors, defined as the difference between the actual and predicted outcomes. The first nine months form the pre-intervention sample, and the ML prediction errors in this period are utilized to estimate inherent biases of the ML prediction, the biases naturally embedded by the use of ML techniques irrelevant to policy interventions; the last nine months correspond to the post-intervention sample during which the ML model estimates counterfactual outcomes for the pilot stocks, the hypothetical outcomes in the absence of TSPP during

[11] The regression approach has been the universal choice of empirical framework among the prior works on TSPP. In the context of RCT, however, it may not be the right choice. For example, "Regression methods were not originally developed for analyzing data from randomized experiments, and the attempts to fit the appropriate analyses into the regression framework requires some subtleties." ((Athey and Imbens, 2017, p. 94)) and "Experiments should be analyzed as experiments, not as observational studies. A simple comparison of rates might be just the right tool, with little value added by 'sophisticated' models" ((Freedman, 2006, Abstract)).

[12] It refers to a cross-sectional characteristics of whether a new tick size of $0.05 likely becomes a binding constraint on quoted spread.

[13] For related discussions, see Athey and Imbens (2017, p.98).

the post-intervention period. Finally, averaging the prediction errors over the pre- and post-intervention periods separately and having another difference in those averages of the two periods, this ML procedure estimates ITEs for a given outcome stock-by-stock with bias correction in place.[14] Similar to Abadie, Diamond, and Hainmueller (2010) and Abadie, Diamond, and Hainmueller (2015), the ML procedure generates the null distribution consisting of the ML estimates for the non-pilot stocks, the stocks not belonging to TSPP. Against it, I perform inference on ITEs, not only for stocks in the treatment groups but also for those in the control group of TSPP one-by-one.

With ATE estimates for each of the three outcomes obtained from standard fixed-effect panel-data regressions as benchmark, I find that while ATEs on percentage quoted spread turn out to be significantly positive, less than half of the sample stocks show significant ITE estimates; ATEs on consolidated displayed depth appear to be significantly positive, which is in line with the overall results of hypothesis testing at the individual level; ATEs on high-low volatility are estimated as significantly negative, but more than 95% of the sample stocks do not show significance. I find that differing results of significance between ATEs and ITEs are attributed to the presence of extreme values of ITEs. Also, a ML estimator of ATE, defined as average of ITE estimates with no involvement of the control group, gives the results statistically indistinguishable from those obtained from panel-data regressions on the RCT design, confirming reliability of the proposed ML approach.

Importantly, I uncover the pilot stocks in the control group rarely show significance at the individual level for all the outcomes, going against the previous findings that document significant spillover effects on a variety of outcomes over the control group (e.g., Chung, Lee, and Rösch (2020); Rindi and Werner (2019)).[15] I find it reasonable given that traders do not have the incentive to quote in nickel tick size when they can quote in penny tick size, leaving little room to the control group stocks getting "treated" in the first place.

In cross-sectional Probit regressions that takes significance of ITEs as the binary outcome for the stocks in the treatment groups, I show that tick constrainedness can predict policy-effect significance only for percentage quoted spread but not for consolidated displayed depth.[16] That is, most of the pilot stocks that receive significant positive effects on percentage quoted spread in the post-intervention periods are those whose average quoted spreads in the pre-intervention periods are smaller than \$0.05, which is consistent with subgroup analysis of Chung, Lee, and Rösch (2020). However, this is not the case for consolidated displayed depth. Increases in it are widely observed regardless of degrees of tick constrainedness. This would reflect pulling-up effects of price-choice restriction imposed under the nickel tick size. That is, for treated pilot stocks that were previously quoted at penny ticks, traders now can only choose multiples of nickel, thereby tending to quote more often at top-of-the-book prices when their valuations are lower than the top-of-the-book prices within a few pennies. Interestingly, even controlling for tick constrainedness, price level, and market capitalization, percentage realized spread, at any choice of 30-sec., 1-min., or 5-min. time horizon, still shows significant predictability on the both outcomes, indicating that less-profitable stocks for liquidity providers in the pre-intervention periods,

[14] This describes a sort of "difference-in-difference" operation. It is designed to get ride of inherent biases of ML predictions, leaving the ML estimates only due to the policy intervention in TSPP. This debiasing strategy is built upon a "fair" comparison in a sense that the predictions errors of the both pre- and post-intervention sessions are all made from the out-of-sample predictions, which is similar to the debiasing scheme suggested in Chernozhukov, Wuthrich, and Zhu (2020) in the context of SCM.

[15] Prior works mainly studied spillover effects in the way of Boehmer, Jones, and Zhang (2020), which is essentially before/after mean comparison of outcome for control group. As widely known, though, this is not necessarily casual comparison because of time effects researchers are hardly able to fully take out even with many control variables in place.

[16] High-low volatility is dropped out of this analysis because of widespread insignificance at the individual level.

gauged by short-run percentage realized spread, are more likely to receive significant policy effects.

The contribution of this paper is largely fourfold. First, this paper complements the previous findings on TSPP that are built upon ATE mostly. Second, this paper introduces a complete ML procedure to a policy evaluation problem from an *empiricist's* perspective, serving as a better user guide for researchers who focus on pragmatism of ML rather than theory. Third, this paper brings a new perspective into ongoing discussion about efficacy of RCTs in policy evaluation. In presence of big data, the paper shows that researchers can study more than ATE in a RCT using ML techniques. Finally, this paper unveils how good ML predictive models can be for ATE-based casual inference. In comparison with RCT estimators of ATE that has been widely accepted as the gold standard, the paper shows that a ML predictive model built upon big data can produce ATE estimates statistically indistinguishable from those of RCT versions.

The remainder of the paper is organized as follows. In Section 2, I briefly introduce TSPP. I summarize previous findings on TSPP and review the literature on tick size. In Section 3, I detail the ML-based empirical approach. In the potential outcome framework, I develop ML estimators of ITE and ATE along with bias-correction and inference strategy. In Section 4, I describe data used in this paper. In this section, I also illustrate the sampling process for pilot stocks. In Section 5, I present empirical results of ML estimation of ITE and ATE. In Section 6, I deliver a policy implication based on the empirical results of ML estimation. Finally, I conclude the paper in Section 7.

2. Tick Size Pilot Program

I start this section with discussing key features of TSPP as a RCT.[17] Then, I go over the existing findings on TSPP with focus on the impact of the increased tick size on market quality measures. Finally, I summarize previous studies in the literature on tick size conducted prior to TSPP.

2.1. Treatments, Randomization, and Outcomes

Two regulatory bodies of the U.S. stock market, Security and Exchange Commission (SEC) and Financial Industry Regulatory Authority (FINRA), launched TSPP in Oct. 2016. TSPP is a RCT that carries small-cap ($3 billion or less) common stocks. The main rule change of interest in TSPP is an increase in the minimum price increment, often called tick size. Under TSPP, roughly 1,200 stocks in three treatment groups commonly face a five-fold increase in quoting tick size from one penny to one nickel over the course of the two-year pilot period. By contrast, the pilots stocks in a similar number in the control group experience no rule change in the same period, serving as a comparison group against the treatment groups.

An increase in tick size is better understood as a restriction on price choice in the decimal price grid. For example, when a stock is traded around $10.05, under the penny tick size can traders make a quote with penny increment, such as $10.01, $10.02, $\cdots$, $10.09, $10.10 for buy or sell. However, they are only able to choose a multiple of nickel, such as $10.00, $10.05, or $10.10, under the nickel tick size.

TSPP also involves two other rule changes, which are applied progressively over the three treatment

[17] There are a number of research papers offering in-depth institutional contexts on TSPP. For example, see, among others, Albuquerque, Song, and Yao (2019), Bartlett and McCrary (2017), Griffith and Roseman (2019), Hu et al. (2018) and Rindi and Werner (2019). Also, SEC discusses them in SEC (2012, p. 2). A general description of TSPP can be found at FINRA's and SEC's TSPP webpages: `https://www.finra.org/rules-guidance/key-topics/tick-size-pilot-program` (FINRA); `https://www.sec.gov/ticksizepilot` (SEC).

groups. The first treatment group ($G1$) receives only the tick size increase in quotation. In addition to it, the second treatment group ($G2$) is enforced to trade in the nickel tick size. Under those two rule changes in place, the last treatment group ($G3$) is also subject to the *trade-at rule* that restricts off-exchange trading.[18]

Importantly, TSPP adopts a stratified random sampling process in assigning pilot stocks to the treatment and control groups. This balances the treatment and control groups on several key cross-sectional characteristics, price, market capitalization, and trading volume, that likely impact on market outcomes regardless of exposure to policy interventions.[19] To be specific, eligible stocks are labeled by either, low, medium, or high on each of the characteristics, resulting in total 27 possible categories.[20] Then, the stocks in each category are randomly assigned to three treatment groups. Each treatment group roughly consists of 400 stocks, and the rest of the eligible stocks not assigned to the treatment groups constitute the control group.

The SEC's online Bulletin letter states that TSPP is designed "to assess whether wider tick sizes enhance the market quality of these [small-cap] stocks for the benefit of issuers and investors—such as less volatility and increased liquidity." It shows market quality centers on outcomes of interest in TSPP. Typical market quality measures include quoted spread, depth, volatility, etc. Furthermore, other types of outcomes can be studied in TSPP as well. For instance, a tick size change apparently impacts on quoted spread, which in turn can affect decision making of traders on order choice (e.g., Bloomfield, O'hara, and Saar (2005); Griffiths et al. (2000); Harris and Hasbrouck (1996); Hollifield, Miller, and Sandås (2004); Ranaldo (2004)) and venue choice (e.g., Buti, Rindi, and Werner (2011); Kye and Mizrach (2019); Ready (2014)). Then, any measure approximating those activities can be also outcomes of interest in TSPP.

2.2. Existing Findings

Reflecting the first-order question of how TSPP impacts on market quality, early empirical investigations into TSPP are putting a variety of liquidity measures on the left hand side and looking to estimate average effects of the TSPP intervention. By and large, the five-fold increase in tick size turns out to be widening quoted spread but, at the same time, ramping up consolidated displayed depth at the National Best Bid and Offers (NBBO) level (e.g., Albuquerque, Song, and Yao (2019); Chung, Lee, and Rösch (2020); Hansen et al. (2017); Hu et al. (2018); Lin, Swan, and Mollica (2018); Penalva and Tapia (2017); Rindi and Werner (2019)). This is consistent with the general understanding of the literature that widening tick size promotes liquidity provision, represented by increases in consolidated displayed depth, at a higher expense of taking liquidity, shown by increases in quoted spread. While the magnitude of the impacts vary from paper to paper, the findings of the prior works overall are significant both statistically and economically. Interestingly, Chung, Lee, and Rösch (2020) and Rindi and Werner (2019) show the presence of the spillover effects in TSPP that the pilot stocks in the control group experienced a certain

[18] For detail, visit https://www.finra.org/rules-guidance/key-topics/tick-size-pilot-program.

[19] Randomization removes only *in expectation* selection biases coming from preexisting differences of outcomes between treatment and control groups. So it could result in a biased estimate in a *given* experiment, as randomization itself does not achieve balance between the treatment and control groups. Stratified random sampling is often recommended in RCT for balancing, thereby generating the treatment and control groups *in practice*, rather than in expectation, reasonably identical except exposure to policy interventions at the group level.

[20] There are certain eligibility conditions for stocks to be included in TSPP. For one, stocks must have market capitalization of $3 billion or less and closing price of at least $2.00 per share, based on September 2, 2016. For full description, see the "SEC's Plan to Implement Tick Size Pilot Program," available at https://www.sec.gov/rules/sro/nms/2015/34-74892-exa.pdf.

degree of "treatment" in the same direction those in the treatment groups went through.

Yet, the results on liquidity leave some ambiguity in deciding the total effect of the tick size increase on liquidity, as both quoted spread and displayed depth at NBBO increase. In reflection of it, Chung, Lee, and Rösch (2020) and Griffith and Roseman (2019) take as outcomes hypothetical multi-share round-trip transaction costs that jointly consider both of the negative and positive aspects of the tick-size effects on liquidity. While Griffith and Roseman (2019) find a detrimental total effect with a Nasdaq-listed stocks sample, Chung, Lee, and Rösch (2020) tell otherwise by showing a decline in round-trip costs, computed from the consolidated limit order book covering all the pilot stocks.

In addition, the impact of the tick-size increase on liquidity likely goes beyond the NBBO level. In this regard, Penalva and Tapia (2017) look at cumulative depths for all the pilot stocks with ITCH data and find increases in depth are not widely observed within a few extra tick ranges. With Nasdaq-listed pilot stocks in ITCH data, Griffith and Roseman (2019) show a reduction in cumulative depth. By contrast, Chung, Lee, and Rösch (2020) document a significant increase in cumulative depth based on more comprehensive data sets from the Thomson Reuters Tick History database.

Other popular liquidity measures are also put in place of outcomes. Overall, effective spread, realized spread, and price impact all appear to go up by the tick-size intervention (e.g., Bartlett and McCrary (2017); Chung, Lee, and Rösch (2020); Hu et al. (2018); Lin, Swan, and Mollica (2018); Penalva and Tapia (2017); Rindi and Werner (2019)). Market efficiency, typically gauged by variance ratios or auto-correlation of short-term returns, tends to deteriorate (e.g., Albuquerque, Song, and Yao (2019); Hu et al. (2018)). Results on volatility are mixed though. While Penalva and Tapia (2017) show a decrease in volatility by the tick-size increase, Hu et al. (2018) and Rindi and Werner (2019) find the opposite. Different ways of approximating volatility may be contributing to those conflicting findings.

Importantly, a majority of the papers look into heterogeneous effects of the tick size increase. They focus on the fact that there is a group of the pilot stocks whose quoted spreads are often less than \$0.05 in the pre-intervention period, and the others tending to have a larger quoted spread, for instance, exceeding \$0.05 in the same period. This observation naturally but logically leads to a thought experiment that switching over to the nickel tick size from the penny likely brings about differing impacts between the tick-constrained and tick-unconstrained groups.[21] In short, they find tick-constrained stocks generally have more pronounced effects on various liquidity measures than tick-unconstrained stocks do mostly in the same direction (e.g., Albuquerque, Song, and Yao (2019); Chung, Lee, and Rösch (2020); Hu et al. (2018); Lin, Swan, and Mollica (2018); Rindi and Werner (2019)).

There are other papers on TSPP that deal with research questions not directly related with the market quality aspect. Mostly, those papers are interested in shedding light on the proliferation of a wide range of trading venues. In particular, researchers employ the tick size change to see the role of diverse fee/rebate schedules or unravel motivation behind off-exchange trading (e.g., Bartlett and McCrary (2017); Cox, Van Ness, and Van Ness (2019); Comerton-Forde, Grégoire, and Zhong (2019); Farley, Kelley, and Puckett (2018); Lin, Swan, and Mollica (2018)). Finally, there are a few studies that look into TSPP to address issues lying in the intersection between market microstructure and corporate finance (e.g., Lee and Watts (2018); Li, Ye, and Zheng (2019); Thomas, Zhang, and Zhu (2018); Ye, Zheng, and Zhu (2019)).

[21] Tick-constrained stocks are those having a narrower quoted spread in the pre-intervention period so that the nickel tick size in the post-intervention period likely becomes a binding constraint on quoted spread; tick-unconstrained stocks are those having a quoted spread wider enough in the pre-intervention period to get the new tick size in the post-intervention period not to be binding quoted spread.

2.3. Literature on Tick Size

Researchers in the market microstructure have long been interested in understanding how tick size affects liquidity.[22] A change in tick size impacts directly on quoted spread that underlies the incentive scheme of liquidity provision and demand (e.g., Goettler, Parlour, and Rajan (2005)). While the magnitude of tick size matters most, the incremental unit of tick size, either fractional or decimal, is also an important consideration in relating tick size and liquidity (e.g., Harris (1997); Harris (1999)). In general, a larger tick size is believed to incentivize liquidity provision while increasing transactions costs of taking liquidity, and vice-versa, even as the details of this presumable effect likely varies by investors' characteristics (e.g., Seppi (1997)), or the composition of informed and liquidity traders in the market (e.g., Anshuman and Kalay (1998); Foucault, Kadan, and Kandel (2005)). Equally important, the fact that liquidity has close ties to price efficiency suggests that a tick-size change is also influential on it, often examined through predictability of short-term returns (e.g., Chordia, Roll, and Subrahmanyam (2008)).

Popular market quality measures, such as quoted spread, realized spread, effective spreads, and displayed depth, and certain volatility measures, are normally put in place of outcomes of interest to evaluate the impact of tick-size changes. While in different marketplaces, most of the tick size changes in the prior literature are driven by either a simple reduction in tick size, from one-eighth to one-sixteenth, or a reduction due to decimalization, from one-sixteenth to one-hundredth (e.g., Ahn, Cao, and Choe (1996); Ahn, Cao, and Choe (1998); Bacidore (1997); Bacidore, Battalio, and Jennings (2003); Bessembinder (2003); Chakravarty, Wood, and Van Ness (2004); Chakravarty, Panchapagesan, and Wood (2005); Chung and Chuwonganant (2002) Goldstein and Kavajecz (2000); Harris (1996); Ronen and Weaver (2001); Van Ness, Van Ness, and Pruitt (2000)). Importantly, all the previous empirical evidence have been collected from certain events of tick size decrease in the form of natural experiments, in contrast to TSPP focused on an increase in the tick size in a RCT.

Yet, the impacts of tick size changes on market quality in general are not uniquely signed nor uniform in the cross-section. Among the sources of the cross-sectional heterogeneity is price level of stocks, to which relative tick size is differently imputed into cost/benefit calculations of order submission strategies (e.g., Harris (1994)). In addition, heavily traded stocks, often classified as large cap stocks, experiences more pronounced effects on liquidity in response to a tick size change (e.g., Bessembinder (2003); Chung and Chuwonganant (2002)). Degree of inter-market competition adds another channel (e.g., Ahn, Cao, and Choe (1996)). Moreover, tick sizes appear to play a pivotal role in other dimensions as well, such as stock splitting (e.g., Angel (1997)), the performance of mutual funds (e.g., Bollen and Busse (2006)), off-exchange routing decisions (e.g., Kwan, Masulis, and McInish (2015)), and IPO decisions (e.g., Bessembinder, Hao, and Zheng (2015)).

3. Machine Learning Approach

This section introduces a ML procedure to investigate policy effects individually. The outline of the ML strategy is the following: (a) build a ML prediction model for a single pilot stock; (b) predict counterfactual outcome for the stock over the sample period; (c) estimate its policy effect as the time-series mean of the differences between actual and predicted outcomes; (d) apply the same procedure

[22] An overarching summary of the prior literature along with related institutional, regulatory backgrounds is available at SEC (2012).

to the rest of the pilot stocks in parallel way. In development of this ML procedure, I discuss a bias-reduction strategy and inferential scheme. In addition, I suggest a ML estimator of panel-data ATE as average of ML estimates of ITE that dose not count on the control group of TSPP at any stage. This is to examine how well a ML-based predictive model is performing in contest with panel-data regression models that exploit the RCT design.

3.1. Cross-sectional Universe and Sample Period

The ML strategy of this paper employs stocks not only in TSPP but those outside TSPP. I call the group of selected stocks outside TSPP a donor pool, which represents *TSPP-free* stocks.[23] In turn, I consider any group in TSPP, either treatment or control group, as a "treatment" group that is subject to direct or indirect influence of TSPP; that is, the cross-sectional dichotomy for casual inference is pilot stocks versus non-pilot stocks.

In constructing a ML model for pilot stock i, there is no involvement of the rest of the pilot stocks including those in the control group but only the non-pilot stocks in the donor pool will be present in parallel manner as they do to the ML model for pilot stock j. Thus, I will focus on the buildup of a ML model for one prototypical pilot stock.

[Figure 3 about here]

I denote the entire cross-sectional universe of the ML model for pilot stock k by set $I \equiv \{k\} \cup I_d$, where I_d represents the donor pool of N_d non-pilot stocks. Accordingly, the empirical setting starts with $(N_d + 1) \times T$ panel data with outcome $Y_{i,t}$ for stock $i \in I$ in period $t \in P$, where $P \equiv \{1, 2, \cdots, T\}$ indicates the whole sample period, consisting of training sample P_{tr}, pre-intervention sample P_{pre}, and post-intervention sample P_{post} with T_{tr}, T_{pre}, and T_{post} periods, respectively, under $T_{tr} + T_{pre} + T_{post} = T$. As Figure 3 illustrates, P_{tr} covers the first 12 months of the whole sample period, Jan. - Dec., 2015; P_{pre} the next 9 months, Jan. - Sept. 2016; P_{post} the last nine months, Nov. 2016 - Jul. 2017.[24]

3.2. Potential Outcome Framework

The ML strategy for estimating ITE is based on the potential outcome framework or the Rubin Causal Model.[25] For pilot stock k, it postulates two potential outcomes $Y_{k,t}(0)$ and $Y_{k,t}(1)$ in period $t \in P$ that represent the outcomes implied with and without policy intervention of interest, respectively. Then, $\Delta_{k,t} \equiv Y_{k,t}(1) - Y_{k,t}(0)$ for $t \in P_{post}$ defines a policy effects in period t. Notice, though, that researchers cannot observe counterfactual, $Y_{k,t}(0)$ in $t \in P_{post}$, whereas $\Delta_{k,t}$ needs knowledge of it.[26]

To estimate counterfactual, I exploit a cross-sectional predictive relation between pilot stock k and N_d non-pilot stocks, similar to the identification strategy of SCM. A rationale behind it is that they are likewise the entities of the U.S. stock market having been interacted under the same market structure.[27] There are thousands of non-pilot stocks that are likely to have a stable cross-sectional relation with pilot

[23] Composition of a donor pool will differ by outcomes. Construction of donor pools for a given outcome will be discussed later in this section and related statistics will be present in Section 2.4.3.

[24] Notice that the policy interventions in TSPP begin in three stages over Oct. 2016. To avoid complications occurring in that month, I delete this month from the sample period, adding one month gap between the pre- and post-intervention periods.

[25] For detail of the potential outcome framework, see Athey and Imbens (2017) and Imbens and Rubin (2015).

[26] This refers to the so-called *fundamental problem of the causal inference*, coined by Holland (1986, p. 947).

[27] Statistically speaking, this would indicate a common-factor model, which is one of the models presumed in SCM.

stock k as an individual or group but at the same time unlikely to be affected by TSPP. To best take advantage of this presumable cross-sectional predictive relation, I let a ML algorithm choose among them in a way of maximizing predictive accuracy for counterfactual estimation.

To bring non-pilot stocks into the potential outcome framework, let $W_{i,t} \in \{0,1\}$ be the binary indicator showing the policy intervention status for stock $i \in I$ in period $t \in P$ as follows:

$$W_{i,t} = \begin{cases} 1 & \text{if } i \notin I_d \text{ and } t \in P_{post}, \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Then, the observable outcomes, denoted by $Y_{i,t}^{obs}$, can be written as:

$$Y_{i,t}^{obs} = Y_{i,t}(W_{i,t}) = \begin{cases} Y_{i,t}(1) & \text{if } W_{i,t} = 1, \\ Y_{i,t}(0) & \text{if } W_{i,t} = 0. \end{cases} \tag{2}$$

To clarify the prediction problem concerned here, further define $\mathbf{Y}_{d,tr}^{obs}$, $\mathbf{Y}_{d,pre}^{obs}$, and $\mathbf{Y}_{d,post}^{obs}$ as $N_d \times T_{tr}$, $N_d \times T_{pre}$, and $N_d \times T_{post}$ matrices of the observed outcomes for the stocks in donor pool I_d, blocked by P_{tr}, P_{pre}, and P_{post}, respectively, i.e., the (i,t)-th entry of each matrix represents $Y_{i,t}^{obs}$. Similarly, $\mathbf{Y}_{0,tr}^{obs}$, $\mathbf{Y}_{0,pre}^{obs}$, and $\mathbf{Y}_{0,post}^{obs}$ denote $1 \times T_{tr}$, $1 \times T_{pre}$, and $1 \times T_{post}$ row-vectors of the outcomes for the single pilot stock over the three respective time blocks. Then, all the observed outcomes indexed by $(i,t) \in I \times P$, denoted by $\mathbf{Y}^{obs}$, can be written as:

$$\underset{(N_d+1)\times T}{\mathbf{Y}^{obs}} = \begin{pmatrix} \mathbf{Y}_{0,tr}^{obs} & \mathbf{Y}_{0,pre}^{obs} & \mathbf{Y}_{0,post}^{obs} \\ \mathbf{Y}_{d,tr}^{obs} & \mathbf{Y}_{d,pre}^{obs} & \mathbf{Y}_{d,post}^{obs} \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_{0,tr}(0) & \mathbf{Y}_{0,pre}(0) & \mathbf{Y}_{0,post}(1) \\ \mathbf{Y}_{d,tr}(0) & \mathbf{Y}_{d,pre}(0) & \mathbf{Y}_{d,post}(0) \end{pmatrix} \tag{3}$$

Recall that the policy effect in each period depends on both $\mathbf{Y}_{0,post}(1)$ and $\mathbf{Y}_{0,post}(0)$, as discussed earlier, although only is the former observable. The statistical problem faced with this policy problem is then to impute the unobserved part using all the available information:

$$\underset{(N_d+1)\times T}{\mathbf{Y}(0)} = \begin{pmatrix} \mathbf{Y}_{0,tr}(0) & \mathbf{Y}_{0,pre}(0) & ? \\ \mathbf{Y}_{d,tr}(0) & \mathbf{Y}_{d,pre}(0) & \mathbf{Y}_{d,post}(0) \end{pmatrix} \tag{4}$$

This has a similar structure with Doudchenko and Imbens (2016, p. 4) that extends the SCM framework to a big data setting. However, there is one important distinction that I reserve one extra set of the unaffected periods, P_{pre}. Those are the periods prior to the beginning of policy intervention but not part of the training sample. P_{pre} plays an important role in this ML procedure. I use it to estimate inherit biases of ML predictions, which is later to be employed to correct the counterpart over P_{post}.[28]

3.3. Estimation of Individual Stock-Level Treatment Effect (ITE)

To impute the unobserved potential outcome in (**??**) and so estimate ITE after all, I set up a Elastic Net regression model.[29] The ML estimation for pilot stock k involves three steps. The first step performs training the ML model using data in P_{tr}; the second step calculates inherent ML biases of out-of-sample

[28] A similar idea of bias correction in the SCM context is recently formulated in Chernozhukov, Wuthrich, and Zhu (2020).

[29] I employ **glmnet** package in `R` for estimation. It is among the most popular ML programming packages that offer a complete set of ML analytic tools, covering from cross-validation to out-of-sample predictions.

predictions over P_{pre}; the final step estimates ITE over P_{post} with differencing out the ML biases obtained from the prior step.

To be specific, training the ML model for stock k with data measured at the half-hour frequency in P_{tr} is based on the following penalized least squares problem: for a given pair of $\lambda > 0$ and $\alpha \in (0,1)$, I solve

$$\begin{aligned}
&\min_{(\mu,w)\in\mathbb{R}\times\mathbb{R}^L} Q(\mu,w|P_{tr};\lambda,\alpha),\\
&Q(\mu,w|P_{tr};\lambda,\alpha) \equiv \sum_{t\in P_{tr}} \left(Y_{k,t}^{obs} - \mu - wX_t'\right)^2 + \lambda\Big(\frac{1-\alpha}{2}\sum_{l=1}^{L} w_l^2 + \alpha\sum_{l=1}^{L}|w_l|\Big), \qquad (5)\\
&X_t \equiv [Y_{d,t}^{obs\prime} \vdots Z_t']', \quad Y_{d,t}^{obs} \equiv [Y_{1,t}^{obs}, Y_{2,t}^{obs}, \cdots, Y_{j,t}^{obs}, \cdots Y_{N_d,t}^{obs}]'_{j\in I_d}, \quad Z_t \equiv [Z_{1,t}, Z_{2,t}, \cdots, Z_{M,t}]'
\end{aligned}$$

where Z_t is a M-vector that consists of M extra predictors measured in period t, other than the outcomes for non-pilot stocks in I_d; $Y_{d,t}^{obs}$ is a N_d-vector that contains the outcomes for non-pilot stocks in I_d observed in period t; X_t is a L-vector with $L = N_d + M$ whose first block is $Y_{d,t}^{obs}$ and second Z_t; $Y_{k,t}^{obs}$ is the observed outcome for pilot stock k in period t.

Notice that X_t contains not only outcomes for non-pilot stocks, $Y_{d,t}^{obs}$, but also extra predictors Z_t exogenous of TSPP. To it, I include VIX, its highest and lowest values, different levels of time fixed effects converted into dummy variables (13 half-hour intervals, five weekdays, and 12 months fixed effects), and all possible interactions among them, making Z_t a 3,119-vector after all. A anologous strategy is found in Burlig et al. (2017) that set up separate ML models for multiple treated units with the same structure of the predictors, including outcomes for untreated units, a variety of time fixed effects, and their interactions.[30]

Before running the parameter estimation on (**??**), I standardize all the right-hand-side variables so that each predictor in X_t has mean zero and variance one. To determine two tuning parameters λ and α, I conduct the 10-fold cross-validation with the training sample, where I limit the value of α to a finite set $\{0.1, 0.2, \cdots, 0.9\}$ but allow that of λ to cover all possible positive real numbers on $(0,\infty)$ as in Doudchenko and Imbens (2016). In the end, the training process returns the Elastic Net estimates $(\hat{\mu}^{tr}, \hat{w}^{tr})$ such that:

$$(\hat{\mu}^{tr}, \hat{w}^{tr}) = \underset{(\mu,w)\in\mathbb{R}\times\mathbb{R}^L}{\arg\min}\ Q(\mu,w|P_{tr};\lambda_{cv},\alpha_{cv}) \qquad (6)$$

where λ_{cv} and α_{cv} are the tuning parameters drawn from the 10-fold cross validation.

With $(\hat{\mu}^{tr}, \hat{w}^{tr})$ at hand, I perform out-of-sample predictions $\hat{Y}_{k,t}^{ML} \equiv \hat{\mu}^{tr} + \hat{w}^{tr}X_t'$, $t \in P_{pre} \cup P_{post}$ and compute the average prediction errors over P_{pre} and P_{post}, denoted by $\hat{\Delta}_{k,pre}$ and $\hat{\Delta}_{k,post}$, separately, as

[30] They employ a ML technique within the panel-data regression framework. By contrast, this paper uses it rather in the SCM context.

follows:[31]

$$\begin{aligned}
\hat{\Delta}_{k,pre} &\equiv T_{pre}^{-1} \sum_{t \in P_{pre}} \hat{\Delta}_{k,t} \\
\hat{\Delta}_{k,post} &\equiv T_{post}^{-1} \sum_{t \in P_{post}} \hat{\Delta}_{k,t} \\
\hat{\Delta}_{k,t} &\equiv Y_{k,t}^{obs} - \hat{Y}_{k,t}^{ML}, \; t \in P_{pre} \cup P_{post}
\end{aligned} \tag{7}$$

In (7), $\hat{\Delta}_{k,post}$ is the average prediction error on the outcome for pilot stock k over the post-intervention period, presumed to be sum of ITE and a bias component that arises due the regularization of the ML technique. On the other hand, $\hat{\Delta}_{k,pre}$ is the counterpart over the pre-intervention period that contains only the bias component.[32] To wipe out the bias component in $\hat{\Delta}_{k,post}$, thus, I take another difference on the average prediction errors over the pre- and post-intervention periods, i.e., the final estimator of ITE for pilot stock k is defined as:

$$\widehat{ITE}_k \equiv \hat{\Delta}_{k,post} - \hat{\Delta}_{k,pre} \tag{8}$$

Under the stationarity assumption on prediction errors in additivity, this bias-correction strategy would be theoretically valid, leaving to $\widehat{ITE}$ only variations induced by policy intervention of interest.[33] To avoid unequal sample-size effects, I also set the lengths of the two periods, P_{pre} and P_{post}, roughly the same as the nine months.

Finally, I define the ITE test statistics, denoted by $\hat{\tau}_k$, as normalized differences:

$$\hat{\tau}_k \equiv \frac{\hat{\Delta}_{k,post} - \hat{\Delta}_{k,pre}}{\sqrt{\hat{S}_{k,post}^2 + \hat{S}_{k,pre}^2}} \tag{9}$$

where $\hat{S}_{k,\cdot}^2$ is a long-run variance estimator for $\hat{\Delta}_{k,t}$ based on the Newey-West formula with lag truncation parameter $m = 13$ (half-hour intervals) $\times$ 5 (days), in reflection of potential autucorrelation over one trading week.[34]

Statistical inference at the individual stock level with (9) will be performed based on the *placebo test*, popularized by the early SCM literature.[35] Unlike standard econometric approaches, this does not rely on asymptotic distributions. I instead generates the null distribution from ITE results for the non-pilot stocks whose test statistics reflect uncertainty of the ITE estimator under the *null*. In investigating ITE, I will mostly focus on significance of ITE with test statistics (9) rather than the estimator (8) itself.

[31] To relieve outlier effects, I trim out the ML prediction errors $\{\hat{\Delta}_{k,t}\}_{t \in P}$ at the lower and upper 2.5 percentiles over P_{pre} and P_{post} separately.

[32] As widely known, there are naturally embedded biases in ML predictions. This is because typical ML prediction models aim to maximize predictive accuracy by balancing bias square and variance in coefficient estimation, which does not necessarily set bias zero. The Elastic Net is no exception, and will return coefficient estimates $\hat{\mu}$ and $\hat{w}$ that likely allow nonzero biases in exchange for a better out-of-sample predictive accuracy. This is in short called regularization biases

[33] This is a widely accepted assumption in the time-series context. A similar idea is employed in Chernozhukov, Wuthrich, and Zhu (2020) to correct ML bias in the SCM context.

[34] The Newey-West variance formula with truncation parameter m is given by:

$$\hat{S}_{k,\cdot}^2 = \hat{\gamma}_{k,\cdot}(0) + 2 \sum_{r=1}^{m-1} (1 - r/m) \hat{\gamma}_{i,\cdot}(r), \quad \hat{\gamma}_{k,\cdot}(r) = T_{\cdot}^{-1} \sum_{t \in P_{\cdot}} \left(\hat{\Delta}_{k,t} - \hat{\Delta}_{k,\cdot}\right)\left(\hat{\Delta}_{k,t-r} - \hat{\Delta}_{k\cdot}\right)$$

[35] For overview, see Abadie, Diamond, and Hainmueller (2015, pp. 499-500).

3.4. Hypothesis Testing

Hypothesis testing is based on placebo tests. The principle of placebo tests here is that a policy effect for a pilot stock shall be considered as significant only when it is clearly differentiated from those obtained from non-pilot stocks in the same manner.[36]

To be specific, I compute test-statistics in (**??**) for every non-pilot stock $j \in I_d$ following the same procedure but excluding self-inclusion to $Y_{d,t}^{obs}$ in (**??**). After that, I draw the distribution of the test statistics of them $\{\hat{\tau}_j\}_{j \in I_d}$ and find critical values at 5% significance level.[37] Viewing it as the null distribution, I perform hypothesis testing for each of the pilot stocks one-by-one.

As widely known, however, this strategy potentially poses the multiple testing problem, as there are hundreds of individual testings involved.[38] Taking this into account, I apply the *Benjamini–Hochberg* (BH) procedure to each group of TSPP separately and base main individual testing results on it.[39]

3.5. Comparison with Existing Approaches

Most of the prior works focus on ATE, which is concerned with the null hypothesis of zero ATE. This reflects the effort to understand policy effects in *average sense* for the population of interest. Exploiting the RCT design of TSPP, one can also run the randomization inference, testing the sharp null that there is no policy effect at all for all intervened units (e.g., Young (2019)). Unlike to testing on zero ATE, the sharp null approach is interested in the *existence* of policy effects. Different from both, the placebo-test approach of this paper carries the null hypothesis of no policy effect for one given unit. It is designed to evaluate policy effects *unit-by-unit*. In the context of policy evaluation, this would be painting the most comprehensive picture of policy effects.[40]

3.6. Estimation of Average Treatment Effect

I also consider a ML estimator of panel-data ATE, simply defined as average of ITE estimates in (8). This ML estimator of ATE is immediately comparable to those obtained from typical panel-data regression models, such as Difference-in-Difference. Because of comparability, this spin-off ML estimator helps examine how reliable a ML-based predictive model is in comparison with panel-data regression models built upon the RCT design, or the "gold standard."

In TSPP, there are three nonoverlapping treatment groups of pilot stocks, denoted by $G1$, $G2$, and $G3$, to which three separate policy changes are applied progressively from $G1$ through $G3$, as explained in Section 2. In this exercise, I do not try to break apart treatment effects by each change. Instead, I simply

[36] The pioneers of SCM describes it as: "our confidence that a particular synthetic control estimate reflects the impact of the intervention under scrutiny would be severely undermined if we obtained estimated effects of similar or even greater magnitudes in cases where the intervention did not take place" ((Abadie, Diamond, and Hainmueller, 2015, p. 499)).

[37] To avoid outlier effects, I trim out $\{\hat{\tau}_j\}_{j \in I_d}$ at lower and upper 0.5% quantiles before proceeding with statistical inference.

[38] Some of the p-values less than a predetermined significant level like 5% could be driven purely by chance even if all the null hypotheses are indeed true; that is, the probability of committing false positives, i.e., a chance of a null hypothesis being spuriously called significant, gets larger simply because the number of the hypothesis tests involved is large.

[39] An excellent summary of the BH procedure with an example is offered by McDonald (2014, p. 257) or its online version at `http://www.biostathandbook.com/multiplecomparisons.html`. As to be present later, though, the BH procedure does not bring in a dramatic change to the results of inference drawn from regular p-value-based testing.

[40] Another important advantage is that this approach is easy-to-understand compared to the other inferential strategies involving ML techniques for casual inference. One drawback of this approach is that it does not provide confidence intervals of effects. It is, however, of little importance in this paper.

look into ATE group-by-group, which is actually the common way a majority of the prior works have adopted.

Employing the ML estimator of ITE in (8), the ML estimator of panel-data ATE for treatment group G is defined as:

$$\widehat{ATE}_G^{ML} \equiv N_G^{-1} \sum_{i \in G} \widehat{ITE}_i, \quad N_G \equiv \#\{i \in G\}. \tag{10}$$

Accordingly, the standard error and variance estimator of it are defined as:

$$\widehat{SE}_G^{ML} \equiv \sqrt{\frac{\widehat{\mathbb{V}}_G^{ML}}{N_G}}, \quad \widehat{\mathbb{V}}_G^{ML} \equiv (N_G - 1)^{-1} \sum_{i \in G} (\widehat{ITE}_i - \widehat{ATE}_G^{ML})^2. \tag{11}$$

3.7. Construction of Donor Pools

One set of the predictors in the ML model is the outcomes for non-pilot stocks, $Y_{d,t}^{obs}$ in (5), motivated by the SCM literature. In turn, this necessitates a sampling procedure to form a well-designed donor pool of non-pilot stocks.

Abadie, Diamond, and Hainmueller (2015, p. 500) lay out three criteria for units to be part of donor pool in the SCM context. I find two of them directly applicable to this paper, as stated: (a) the units in the donor pool shall not experience significant idiosyncratic shocks to the outcomes of interest; (b) the units in the donor pool shall not be affected by policy intervention of interest.

I construct the donor pool for each outcome according to those two criteria. I apply several filtering rules for (a) at the data processing level and examine stationarity on time series of outcomes for (b). Because of this process, the whole sampling procedure results in different sets of non-pilot stocks depending on choice of outcome among percentage quoted spread, consolidated displayed depth, and high-low volatility. The detail of the sampling procedure for donor pools will be presented in subsection 4.3 and Appendix C.

4. Data

This section delivers data descriptions. First of all, the whole sample period is stretched over total 31 months in Jan. 2015 - Jul. 2017, divided by three consecutive segments: the training sample (Jan. - Dec. 2015), the pre-intervention sample (Jan. - Sept. 2016), and the post-intervention sample (Nov. 2016 - Jul. 2017).[41]

I start this section with introducing data sets used in this paper. Then, I define the three outcomes of interest and describe sampling procedures for pilot stocks as well as non-pilot stocks. Finally, I look at descriptive statistics on the three outcomes.

[41] I exclude the four early closing days from the sample: Nov. 27, 2015, Dec. 24, 2015, Nov. 25, 2016, and Jul. 3, 2017.

4.1. Data Sets

Daily TAQ and CRSP are the two most important data sets in this paper. The former provides quote data for all the U.S.-traded stocks that record intraday updates of best orders on the either bid or offer side in major national stock exchanges. All the three outcomes of interest are constructed from it, as to be shown below. The latter contains information of security-specific characteristics and trading summaries, such as the number of outstanding shares, listing exchanges, stock classes, opening/closing price, and share trading volume, for most of the U.S.-traded stocks on a daily basis. I employ it to sample stocks. In addition, I obtain CBOE Volatility Index (VIX) at the half-hour frequency from the Bloomberg Terminal. Lastly, data tracking the list of pilot stocks over the experimental periods are found at the FINRA's TSPP page.[42]

4.2. Outcomes

The three outcomes of interest describe three different aspects of U.S. stock market quality, consisting of two liquidity measures, percentage quoted spread and consolidated displayed depth, and one short-term volatility measure, high-low volatility. All the outcomes are based on NBBO quotes, the nationwide best quotes among the locally best quotes on individual stock exchanges.[43] Following the related literature, I sort out NBBO at each quote update in the raw TAQ data using the Holden and Jacobsen (HJ) algorithm.[44]

The three outcomes, indexed by stocks and half-hour intervals, are computed in the standard way of the literature. Percentage quoted spread, the national best offer price minus the national best bid price divided by the midpoint of them at each quote update, is time-averaged within the half-hour intervals. Similarly, consolidated displayed depth, the sum of the displayed bid and offer depths on all the exchanges at NBBO divided by two at each update, is also time-averaged within the half-hour minutes. High-low volatility is defined as the highest NBBO midpoint minus the lowest NBBO midpoint in each half-hour interval divided by the time-averaged NBBO midpoint on the corresponding half-hour interval. Finally, I multiply percentage quoted spread and high-low volatility by 10,000 and consolidated displayed depth by 100 to represent percentage quoted spread and high-low volatility in basis points (bps) unit and consolidated displayed in shares unit, respectively.

4.3. Sample Stocks

While all the pilot stocks in TSPP are of interest by default, the use of machine learning approaches requires certain data conditions to ensure existence of reliable training data sets, which excludes some pilot stocks inevitably. As for non-pilot stocks, a sampling procedure, in addition to a similar require-

[42] For detail, visit https://www.finra.org/rules-guidance/key-topics/tick-size-pilot-program.

[43] The U.S. stock market is built on a virtual consolidation among 13 national exchanges. They are standing independently as separate markets, but are loosely connected as one large stock market by regulation. NBBO depicts overall quality of the U.S. stock market as a whole.

[44] This is the algorithm developed in Holden and Jacobsen (2014) and now becomes the standard procedure in the literature. In that paper, the authors originally discuss data problems in Monthly TAQ quote data for NBBO processing. However, many of the data issues, such as crossed or withdrawn quotes, are still present in Daily TAQ quote data when one runs the NBBO processing from scratch using Daily TAQ quote data only. The authors maintain a SAS code that extends their algorithm to make it applicable to Daily TAQ, which is generously available on one of their websites, https://kelley.iu.edu/cholden. Adopting the same filtering rules found in that SAS code, I write my own JAVA code to run it over a large scale of stocks-days jobs.

ment on trading activities, involves certain stationarity conditions on outcomes, excluding those showing irregular time-series movements during the pilot period.

Pilot Stocks

Based on the FINRA data tracing changes in security, I construct a sample of the pilot stocks. First, I exclude pilot stocks that undergo changes irrelevant to TSPP but likely influential on the outcomes.[45] I further put certain requirements on trading activities to ensure that the resulting sample does not include those infrequently traded or quoted.

[Table 2 about here]

Table 2 summarizes the whole process of sampling the pilot stocks. On top of that, the pilot stocks that reveal unrelated changes, such as changes in ticker symbols or listing exchanges, are excluded. Also, I delete the stocks in the treatment groups switching over to the control group in the pilot periods due to stock prices falling below the minimum required level, \$1.00. In addition, I remove those that drop out of the program too early or do not have complete records in the tracking information, `TSPilotChanges.txt`. With the one-year minimum length of survival in the experimental periods imposed, I further require 5,000-share daily trading at least for two-thirds of the trading days in each of the three sub-samples, the training sample, pre-intervention sample, and post-intervention sample. Finally, I exclude several remaining stocks that appear ill-defined in the initial sample periods in CRSP and NBBO-sorted quote data. In sum, the sampling procedure results in 810 stocks in control group and 272, 257, and 242 stocks in the three treatment groups from $G1$ to $G3$ in order. The selected pilot stocks in control group account for 83.72% and those in the treatment groups 85.74% of the aggregate trading volume over the whole sample period, capturing a majority of trading activities during this period.

Donor pools

Table 3 summarizes the sampling process for non-pilot stocks. Starting with all the U.S.-listed stocks uniquely identifiable from CRSP data, I exclude the stocks involved in TSPP first and place the 5,000-share daily trading rule, as applied to pilot stocks, initially producing 5,955 stocks.[46] Next, I delete the stocks likely to undergo idiosyncratic events over the sample periods. The three abnormality filters, no stock split, no excessive overnight return, and no abrupt change in outstanding shares, conduct it, excluding 1,403 in total and leaving 4,552 stocks. Furthermore, I only consider the stocks that have the same trading sequence as at least one sampled pilot stock over the sample periods.[47] This rule removes 2,221 stocks, but they as a whole represent a very small portion of trading activities, taking off merely less than seven percentage points in the trading volume share from the prior stage. After that, I preclude the stocks that have very low volatility over the sample periods of 520 trading days, gauged by standard deviations of daily returns. It takes out the stocks, like Treasury bill ETFs, whose volatility is less than 15 bps, the volatility level at the 10th percentile on this stage. At last with CRSP data, I perform trimming

[45] Prior works on TSPP also perform similar sampling procedures before conducting empirical analyses (e.g., Chung, Lee, and Rösch (2020); Rindi and Werner (2019)).

[46] The U.S.-listed stocks in this paper are defined as the stocks listed one of NYSE, AMEX, NASDAQ, and ARCA exchanges, identifiable via the Center for Research in Security Prices (CRSP) data. Pilot stocks along with the institutional schedules of TSPP can be accessed through the FINRA's website: `https://www.finra.org/filing-reporting/archived-pre-pilot-files`.

[47] It is a bit restrictive but necessary because predictors in the proposed ML model, the outcomes for non-pilot stocks,, must have the identical time indexes for the dependent variable, the outcome for a chosen pilot stock.

by price level and market capitalization based on Sept. 1, 2016, which is the last trading day before the pilot stocks are officially determined by FINRA. It restricts the range of both price level and market capitalization between the lower and upper 1% quantiles, thereby leaving 2,220 stocks that account for about 65% in the aggregate trading volume over the initial stage. These 2,220 stocks are the baseline group sorted out at the data processing level.

While there is no permanent dropout among the baseline stocks, when they are matched to NBBO-sorted quote data, some of the stock-day observations are deleted. The deletions occur partly because I consider only the trading days on which the first valid NBBO quote, the first quote well-defined in the HJ algorithm, of a given stock is lying at latest within one hour from the opening bell.[48] Other reasons for the deletions are due to several mismatches between CRSP and NBBO-sorted quote data and exclusion of early closing trading days.

For each non-pilot stock in the baseline group, I further investigate stationarities on half-hour time series of the outcomes stock-by-stock. The stationarity tests include random-walk (or unit root) tests on the pre-intervention periods, standard t-tests on the long-run mean differences between the pre- and post-intervention periods, and structural break tests on autoregressive model between the two periods. Based on the intuition that there would be no qualitative difference in dynamics of time series between the two periods if a non-pilot stock were not affected by TSPP, this time-series investigation will leave only a set of the stocks showing stable time-series dynamics of the outcomes over the both pre- and post-intervention periods.

After all, the donor pool for each of the three outcomes, percentage quoted spread, consolidated displayed depth, and high-low volatility in order has 1,280, 1,343 and 2,045 member stocks. While this strategy is a bit conservative in giving out eligibility for donor pools, the resulting groups of the non-pilot stocks are still large in number to justify the use of ML approaches. The detail of the stationarity tests are introduced in Appendix C.

Descriptive Statistics

Table 4 offers descriptive statistics of the three outcomes of interest, percentage quoted spread, consolidate displayed depth, and high-low volatility. It shows the sample means and sample standard deviations in parentheses of the outcomes for the stocks sampled following the procedures in Table 2 and Table 3.

[Table 4 about here]

First of all, there are quite small differences for all the outcomes between the treatment and control groups of TSPP in the periods before the policy intervention comes into play. In the post-intervention periods, however, percentage quoted spread and consolidated displayed depth appear to increase for the treatment group relative to their changes for the control group. These results are consistent with the previous findings on TSPP that widely document increases in quoted spread and depth for the stocks in the treatment groups relative to those in the control group. On the other hand, high-low volatility for the treatment group turns out to decrease relative to the control group during the post-intervention periods.

Meanwhile, the donor pools show very similar trends with the control group for all the three outcome though magnitudes of the outcomes differ between them. This is expected to a degree as they both are

[48] For example, if stock i on trading day t has the first valid NBBO quote at 10:31 AM, then I do not consider trading day t for stock i but it does not necessarily exclude stock i from the baseline group as stock i may have many other trading days $s \neq t$ that have the first valid quote within one hour from the opening bell.

not explicitly exposed to policy changes by TSPP. Also, the donor pools tend to have lower percentage quoted spread and high-low volatility, and higher consolidated displayed depth, compared to the pilot stocks. This is reflective of the differing composition of their respective member stocks in terms of market capitalization. Remind that the pilot stocks are by construction all small-cap stocks of market capitalization at largest $3 billion. However, there is no restriction imposed on market capitalization in constructing donor pools, rendering the donor pools likely composed of the larger market-cap stocks relative to the TSPP groups.[49]

5. Empirical Results

This section presents the empirical results gleaned from ML analyses. I first look into ITEs through test statistics in (9). Next, I study ATE using the ML estimator of panel-data ATE in (10) in comparison with the results of standard panel-data regressions.

5.1. Individual Stock-Level Treatment Effects (ITE)

Studying ITEs is based on hypothesis testing stock-by-stock, basically counting the pilot stocks significantly affected by TSPP. Using the individual testing results, I further investigate cross-sectional characteristics explaining policy effect heterogeneity among the stocks.

Null Distributions

The test statistics for the stocks in the donor pools draw the null distributions for inference at the individual stock level. Remind that those stocks are not part of TSPP, delivering the falsification results that reflect natural possibilities the ML estimator can take in the absence of TSPP in the cross-section.

[Figure 4 about here]

Figure 4 shows the null distributions for the three outcomes, percentage quoted spread, consolidated displayed depth, and high-low volatility. For each outcome, the gray bars describe the histograms of the test statistics constructed from their respective donor pools. Notably, the null distributions are all close to normal distributions with mean and variance being the sample means and sample variances of the test statistics, shown by red dashed lines. This would be the evidence supporting that the proposed ML procedure is well-designed, where the resulting test statistics, normalized bias-corrected prediction errors, in the absence of the policy intervention do not show a skewed nor fat-tailed distribution but have symmetric, bell-shaped distributions centered at zero, analogous to limiting distributions that appear in typical econometric models. Similar to the standard way, a pilot stock whose test statistics on a given outcome is relatively large against those null distributions will be judged as the one receiving a significant policy effect on that outcome.

Testing Stocks-by-Stocks

Figure 5, Figure 6, and Figure 7 show the distributions of the ITE test statistics for percentage quoted

[49] It is the stylized fact that the larger market capitalization, the better market quality, such as lower quoted spread, higher depth, and lower volatility, exactly shown for donor pools relatively to the rest of the groups in the table.

spread, consolidated displayed depth, and high-low volatility, respectively, for the sample stocks in the treatment groups. In the figures, the gray bars draw histograms for each treatment group, the blue dashed lines mark the critical values at the 5% significance level, drawn from the null distributions in Figure 4, and the red dashed line indicates cross-sectional averages of the test statistics.

[Figure 5, Figure 6, and Figure 7 about here]

Those figures are concisely portraying how individual stocks are affected by the policy interventions in light of the null hypothesis. Note that the ITE test statistics are normalized estimates of ITE, which would deliver a sense of the distribution of ITEs. For percentage quoted spread in Figure 5, the distributions of the test-statistics between the treatment groups look closely similar from one group to another. In case of consolidated displayed depth, on the other hand, $G3$ in Figure 6 draws a bit different shape of the distribution compared to $G1$ and $G2$. To be specific, $G3$ shows a more clustered distribution around the mean relative to $G1$ and $G2$. This would indicate that there is an additional impact of trade-at-rule beyond the tick size change on consolidated displayed depth, but such extra impacts would not exist on percentage quoted spread. The distributions on high-low volatility for all the groups in Figure 7 describe no significant effect at the individual level due to the policy changes.

[Table 5 about here]

Table 5 shows the numerical results of the individual testing. The $P < .05$ column counts the number of the sample stocks whose p-values, computed from the null distributions, are less than 0.05. The $P_{BH} < .05$ column represents the same testing results but based on the BH procedure with false discovery rate parameter $\alpha = 0.05$. Given that there are hundreds of the individual testings involved, I take the latter as the primary results for the individual testing.

The testing results for percentage quoted spread in Table 5 show that only less than half of the sample stocks reveal significance of policy effects. On the other hand, consolidated displayed depth turn out to be extensively significant, having more than 80% of the sample stocks across the treatment groups show significance at the 5% level. Meanwhile, there are only less than 5% of the sample stocks that receive significant changes on high-low volatility with inconsistent signs to one another. This essentially implies that TSPP does not impact on short-term volatility. Collectively, those show that the tick size increase does not necessarily widen percentage quoted spread but ramp up displayed depth at the NBBO level without incurring extra short-term volatility.

Discussion on Spillover Effects

I examine the issue of spillover effects, the "treatment effects" spilled over to the control group, at the individual stock level. Figure 8 draws the distributions of the test statistics for the pilot stocks in the control group. While there are some that lie in the rejection region at the 5% level, a majority of the ITE test statistics do not show significance for all the three outcomes. The numerical results in Table 5 confirm it, finding no statistical evidence supporting the presence of the spillover effects at the individual stock level.

I find those results reasonable given that traders are unlikely to quote in nickel tick size over penny tick size unless they are enforced to do so. In other words, there is no possible trigger that can lead "treatment" to the control group and so"treatment effect" to the outcomes for the control group. This goes against the findings of the prior works that show the presence of spillover effects at the group level. I suggest two possibilities that they may falsely identify spillover effects. First, a before/after

average comparison of outcome for the control group, adopted by Chung, Lee, and Rösch (2020) and Rindi and Werner (2019), is not necessarily causal comparison because of time effects that, if any, prevail regardless of randomization. For this approach to work properly, researchers must have the full capability of controlling for all the potential time effects irrelevant to TSPP, which is highly unlikely, especially in linear models.[50] Most of all, it is untestable whether possible confounders in the time dimension are fully controlled for.[51] Further, choice of control variables is often made arbitrarily.

Another approach, adopted by Rindi and Werner (2019), is to form a matched sample of non-pilot stocks that are similar to those in the control group to run a group-by-group comparison. By design, however, this approach has a serious identification problem. If there were indeed spillover effects over the control group, it is highly likely that similar non-pilot stocks also share spillover effects because of similarity. That is, non-pilot stocks similar to those in the control group of TSPP on key cross-sectional characteristics would also experience "treatment effects" in the same manner those in the control group of TSPP are affected. This is on the ground that cross-sectional similarity among the pilot stocks between the treatment and control groups is often pointed as the source of the spillover effects (e.g., Boehmer, Jones, and Zhang (2020)).

Policy Effect Heterogeneity

Now I turn to the question of what kinds of individual characteristics can explain differing effects of the policy interventions in the cross-section. The analysis in the prior subsection shows that treatments effects among affected pilot stocks are not uniform. To see what underlies it, I employ as an outcome the results of the individual hypothesis testings and investigate which pre-pilot covariates have predictability of the treatment effect significance over the post-pilot periods. Notice that I exclude high-low volatility here, as there were very few of the stocks showing significance at the individual level for this outcome.

I consider a simple cross-sectional Probit model, in which the binary response variable is whether the ITE test statistics for stock i shows statistical significance at the 5% level, evaluated by $P_{BH} < .05$ in Table 5. For this binary response outcome, I take tick constrainedness, price level, market capitalization, trading volume, and percentage realized spread in the pre-intervention periods as pre-treatment covariates of interest. This choice is an extension of the prior works on TSPP that have solely focused on tick constrainedness. Price level, market capitalization, and trading volume are among the most important cross-sectional characteristics potentially influential on a variety of market quality measures. In particular, those three covariates are used in forming the strata for random assignment of pilot stocks at the design stage, which further justifies the inclusion of them as control variables in this regression model. Percentage realized spread, which measures short-term profits for liquidity provision, is taken to see whether the increasing tick size indeed gets liquidity providers better off, which is presumed to be the pre-condition of liquidity improvement for small-cap stocks in the TSPP context.

The definitions of the covariates here are the following. I define tick-constrained stocks as those whose time-series average of daily time-weighted dollar quoted spread over the pre-intervention periods, Jan. - Sept. 2016, is less than \$0.05 so that the new tick size \$0.05 is likely to become a binding constraint for quoted spread in the post-intervention periods. Percentage realized spread for stock i is the time-series average of daily volume-weighted percentage realized spread over the pre-intervention periods.[52]

[50] Note all the prior works employ linear models, i.e. they also implicitly assumes a linear structure on time effects, which does not have any theoretical, empirical ground.

[51] In fact, the difficulty in drawing causal inference from before/after comparison is one of the main reasons relying on RCT.

[52] I consider three time horizons, 30 seconds, one minute, and five minutes, for percentage realized spread, which are all

Price and market capitalization for stock i are based on the time-series averages of the daily values of opening price over the pre-intervention periods. Similarly, trading volume is the time series average of daily trading volume in the pre-intervention period. Notably, all those cross-sectional characteristics on the right hand side of the Probit model are computed only using the information in the pre-intervention period while the binary response outcome reflects changes of outcomes made in the post-intervention period.

[Table 8 about here]

Table 8 shows the results of the Probit regressions. On top of that, tick constrainedness appears to explain only significant effects for percentage quoted spread but not for consolidated displayed depth. This goes against the previous findings that document heterogeneity with respect to tick constrainedness on a variety of market quality measures including both quoted spread and depth but is consistent with graphical findings in the previous section. Recall that Figure 6 shows the increases in consolidated displayed depth are widely observed across the pilot stocks in the treatment groups. A part of the results would reflect pulling-up effects of price-choice restriction imposed under the nickel tick size. Traders, who was able to freely choose prices inferior than NBBO by a few pennies before TSPP, are now enforced to choose only multiples of nickel under TSPP. When their valuations are lower than NBBO within a few pennies in nickel tick size, thus, it is highly likely that they quote at NBBO rather than doing so at next nickel ticks, which in turn increases depth at NBBO even if the new tick size is not binding quoted spread.

On the other hand, percentage realized spread turns out to explain significant effects for both percentage quoted spread and consolidated displayed depth. Its accountability survives even controlling for tick constrainedness, market capitalization, price level and trading volume. This delivers that the lower percentage realized spread, the more likely is the null hypothesis rejected in the cross-section, indicating that less-profitable stocks for liquidity providers in the pre-intervention periods are more likely to receive significant effects in the post-intervention periods. Those results, robust over choice of different short-term time horizons, would support the basic idea behind TSPP that widening tick size would incentivize liquidity provision for small-cap stocks through a higher margin.

5.2. Average Treatment Effects

The ML estimator in (10) measures panel-data ATE, which is readily comparable to the ATE estimates obtained from standard panel-data regressions. Further, using a graphical approach, I look into cross-sectional ATE over time to see evolution of policy effects during the sample periods. This approach is particularly useful in breaking down panel-data ATE into the cross-sectional and time-series dimensions.

ML Approach vs. Regression Approach

As benchmark estimates of panel-data ATE, I first estimate ATE using panel-data regressions, which has been de facto the only empirical approach of the literature on TSPP, and compare estimation results

time-averaged daily. As usual, percentage realized spread with time horizon h for a trade made at time τ is defined as:

$$\text{Percentage Realized Spread}_\tau = \frac{2D_\tau(P_\tau - M_{\tau+h})}{M_\tau} \times V_\tau$$

where D_τ is the Lee and Ready (1991) buy-sell indicator; P_τ is the transaction price; M_τ is the midpoint of NBBO prevailing at time τ; V_τ is the volume-weight of this transaction over the total daily trading shares.

with those of the ML approach.[53] To this end, I consider a fixed-effect panel-data model (12) and DiD model (13) as follows:

$$Y_{i,t} = \beta_1 G1_i \times Pilot_t + \beta_2 G2_i \times Pilot_t + \beta_3 G3_i \times Pilot_t + \alpha_i + \gamma_t + \epsilon_{i,t} \tag{12}$$

$$Y_{i,t} = \beta_1 G1_i \times Pilot_t + \beta_2 G2_i \times Pilot_t + \beta_3 G3_i \times Pilot_t + \beta_4 Pilot_t + \alpha_1 G1_i + \alpha_2 G2_i + \alpha_3 G3_i + X'_{it}\gamma + \epsilon_{i,t} \tag{13}$$

where $Y_{i,t}$ is an outcome for stock i in half-hour segment t; $Gk_i, k = 1,2,3$ is the treatment group indicator that takes one if stock i belongs to treatment group Gk and zero otherwise; $Pilot_t$ is the treatment period indicator that takes one if period t is in the treatment period and zero otherwise; α_i and γ_t are stocks and periods fixed effects; X_{it} is a set of covariates that control for preexisting differences of the outcome, if any, across treatment and control groups. For DiD model (12), I consider three specifications according to three different sets of controls that have been widely chosen among the prior works: no covariate, VIX, and VIX and log capitalization. In all panel-data models, I adjust the standard errors clustered by both stocks and days.[54]

[Table 6 about here]

Table 6 shows the regression results by the four specifications of panel-data regressions for each outcome, where (M1) is the fixed effect model and (M2), (M3), and (M4) are DiD models with different choice of controls. In the first place, it is worth noting that the main estimates – the coefficients of $G1_i \times Pilot_t$, $G2_i \times Pilot_t$, and $G3_i \times Pilot_t$ – are essentially the same across all the specifications. To a degree, this proves TSPP is a well-conducted RCT, supported by the ATE estimates robust against model specifications.[55]

Since there is no meaningful difference on the estimates of the parameters of interest across the specifications, I only focus on (M1) to investigate ATE in the regression framework. The policy changes in TSPP on average cause 17.77 bps, 14.69 bps, and 14.68 bps increases in percentage quoted spread for treatment group $G1$, $G2$, and $G3$, respectively. For consolidated displayed depth, there are about 2,154-, 2,075-, and 3,052-share increases for $G1$, $G2$, and $G3$, respectively. Finally, the three treatment groups $G1$, $G2$, and $G3$ in order show 9.38 bps, 8.55 bps, and 6.41 bps decreases in high-low volatility. Those are ATE estimates measured relative to the control group.

[Table 7 about here]

For comparison, Table 7 presents the ML and panel-data regression estimates side-by-side, where the regression results are drawn from (M1) in Table 6, which is the fixed-effect panel-data model with only the treatment indicator dummy variables added to the unit and time fixed effects.[56] In short, they are very close to one another for all the three outcomes across all the three treatment groups. The t-values on the differences between them indicate that they are statistically indistinguishable.[57] In sum, it follows

[53] Both ML and panel-data regression approaches estimate ATE over the pre- and post-intervention periods, the nine consecutive months before and after one-month policy phase-in of Oct. 2016.

[54] For panel data regressions, I employ `reghdfe` command in `Stata16`.

[55] A RCT produces an unbiased estimate of ATE *in expectation*, which does not necessarily imply an unbiased estimate of ATE *in practice*. It is possible even under randomization that there is unbalance between treatment and control groups and that an estimate of ATE is contaminated by the preexisting difference between them. In such a case, estimates of ATE can be sensitive to choice of control variables even under proper randomization. For related discussion, see Deaton and Cartwright (2018, p. 4).

[56] Notice that this fixed-effect model nests most of the DiD models adopted in the prior works.

[57] Alternatively, we can see this from the observation that the ML estimates of ATEs are mostly lying within the 95% confi-

that when it comes to estimation of ATEs in panel data, the proposed ML approach is as good as the regression approach, where the former exploits ML predictions to approximate counterfactual while the latter takes advantages of the RCT design for it. Those results might support the view of Hal R. Varian, quoted earlier as, "A good predictive model can be better than a randomly chosen control group, which is usually thought to be the gold standard ((Varian, 2014, p. 24))." This also would be indirect evidence supporting credibility of the proposed ML approach.

Overall, the results for percentage quoted spread and consolidated displayed depth with the half-hour data are consistent with those of a majority of the previous works that investigate similar outcomes with daily versions of data (e.g., Albuquerque, Song, and Yao (2019); Chung, Lee, and Rösch (2020); Hansen et al. (2017); Hu et al. (2018); Lin, Swan, and Mollica (2018); Penalva and Tapia (2017); Rindi and Werner (2019)). For volatility, though, the direction of the average policy effect is the same as Penalva and Tapia (2017) that uses a close measure of short-term volatility in intraday intervals but the opposite to Hu et al. (2018) and Rindi and Werner (2019) that approximate volatility over the full day range. The mixed results on volatility would be partly due to the disagreeing ways of defining volatility among different papers.

Time-Series Analyses

Now I try to break down the ATE results of panel data into the cross-sectional and time-series dimensions using a graphical approach. I look at cross-sectional ATEs day-by-day to trace the evolution of average policy effects for each treatment group over the sample period. This is to see how the ATE estimates in panel data are obtained as those are essentially the pooled mean differences in the cross-sectional and time-series dimensions.

To that end, I define the cross-sectional ATE estimator of RCT for period t as:

$$\widehat{ATE}^{RCT}_{G,t} \equiv N^{-1}_{G,t} \sum_{i \in G} Y_{i,t} - N^{-1}_{c,t} \sum_{i \in C} Y_{i,t} \tag{14}$$

where $Y_{i,t}$ is an outcome for stock i in period t; G and C indicate the treatment and control groups with the number of member stocks $N_{G,t}$ and $N_{C,t}$, respectively, in period t. The variance estimator, as the Neyman variance estimator (see (Athey and Imbens, 2017, p. 89)), is given as:

$$\widehat{V}^{RCT}_{G,t} \equiv \hat{S}^2_{G,t}/N_{G,t} + \hat{S}^2_{C,t}/N_{C,t} \tag{15}$$

where $\hat{S}^2_{G,t}$ and $\hat{S}^2_{C,t}$ are the sample variances of outcome $Y_{i,t}$ in period t for the control and treatment groups, respectively.

As a reference, I also consider a ML version of the cross-sectional ATE in a similar manner:

$$\widetilde{ATE}^{ML}_{G,t} \equiv N^{-1}_{G,t} \sum_{i \in G} \hat{\Delta}_{i,t}, \quad \hat{\Delta}_{i,t} \equiv Y^{obs}_{i,t} - \hat{Y}^{ML}_{i,t} \tag{16}$$

where $Y^{obs}_{i,t}$ and $\hat{Y}^{ML}_{i,t}$ are observed and predicted outcomes, respectively. Notice that unlike the ML estimator of panel-data ATE in (10), $\widetilde{ATE}^{ML}_{G,t}$ here is not bias-corrected.

Figure 9, Figure 10, and Figure 11 draw the daily time-series of the cross-sectional ATE estimates of

dence interval of the estimates of the panel-data regressions.

percentage quoted spread, consolidated displayed depth, and high-low volatility, respectively, for RCT and ML versions.[58] The dotted gray lines draw the 95% confidence intervals of the RCT estimator, computed day-by-day. The shaded regions represent the one-month policy phase-in period, dividing the pre- and post-treatment periods.

[Figure 9, Figure 10, and Figure 11 about here]

First of all, for the all outcomes across the three groups, there is little aspect of heterogeneity in the time dimension. While there are some fluctuations, the time-series of the cross-sectional ATE estimates from RCT appear to be quite stable over time. Those results are expected given that the policy changes considered in TSPP are those immediately influential time-independently once they start to kick in.

Interestingly, the ML estimator is performing as good as the RCT one despite the fact that it is likely to be biased in a statistical sense. The ML estimates are mostly lying within the 95% confidence intervals in the both pre- and post-intervention periods. Those results indirectly explain why the ML approach to estimation of ATE in the panel data turns out be as good as regression approaches, such as DiD, even as it does not count on the control group.

[Figure 12, Figure 13, and Figure 14 about here]

Exploiting such a ML performance, I also draw quantile values of $\{\hat{\Delta}_{i,t}\}_{i\in G}$ day-by-day to project cross-sectional distributions onto the time dimension. Investigating cross-sectional quantiles over time is one way of summarizing policy effects in both the cross-sectional and time-series dimensions.

Figure 12, Figure 13, and Figure 14 show the daily time-series of the quantile values at 20%, 40%, 60%, and 80%, along with cross-sectional averages for percentage quoted spread, consolidated displayed depth, and high-low volatility, respectively. For percentage quoted spread in Figure 12, the times series of the cross-sectional ATE estimates turns out to be a little skewed toward upper values, as it closely follows that of the quantile values at 60%. On the other hand, the graph for consolidated displayed depth in Figure 13 shows that the cross-sectional ATE estimates are mostly driven by large values. The ATE estimates are very close to the quantile values at 80% for the most of the days. Lastly, the trajectory of the ATE estimates for high-low volatility in Figure 14 are almost centered in the cross-section. The time-series of the cross-sectional ATE estimates are lying between the quantile values at 40% and 60% all the times. Summing up, those graphical analyses reveal that the panel-data estimates of ATEs are not necessarily average values under symmetry. Skewness is present, which requires researchers to use caution when taking the ATE estimates in panel data as representative policy-effect metrics for policy learning.

5.3. ATE vs. ITE

So far, I look into ITE and ATE separately. Interestingly, there is some discrepancy of policy effects described by ITE and ATE. Most dramatic, the ATE results on high-low volatility show a negative significant impact, but individual hypothesis testing results find essentially no effect on it. In particular, Deaton and Cartwright (2018) and Young (2019) recently point out vulnerability of statistical inference on ATE in a RCT in the presence of asymmetric ITEs, or outliers. Differing pictures drawn by ATE and ITE on high-low volatility could be subject to this problem.

[58] Because the daily frequency is the one readily interpretable in the time-series dimension, I use daily Y and $\hat{\Delta}$ in this analyses for (14) and (16), respectively, by averaging half-hour values of them on each day per stock.

To see this, I run fixed-effect panel-data regressions of model (12) under trimming performed at the stock level. Notice that ML estimates of ITE, $\widehat{ITE}_i$ in (8), enable to identify which pilot stocks have extreme values. If the ATE estimates from the panel-data regressions are not driven by extreme values, then the estimation results shall not be severely affected by trimming. For each outcome, I consider four cases that exclude from the regression data stocks whose ITE estimates are lying outside (1% , 99%) , (2.5%, 97.5%), (5%, 95%), and (10%, 90%). Table 9 shows the relevant regression results.

[Table 9 about here]

It turns out that dropping the stocks of extreme values brings about quite large changes in coefficient estimates. For percentage quoted spread and consolidated displayed depth, exclusion of the stocks at the lower and upper 5% halves the coefficient estimates for all the treatment groups, as shown in column (4) in the table, though those changes do not alter hypothesis testing results. As for high-low volatility in the same column, however, exclusion of the extreme stocks washes out statistical significance. Those results support the view of Deaton and Cartwright (2018) and Young (2019) discussed above. Once extreme stocks are removed, the discrepancy of policy effects described between ITE and ATE disappears.

6. Policy Implications

Tick size increase incurs a wealth transfer from liquidity demanders to suppliers. A coarser price grid likely enables liquidity suppliers to collect a higher margin from market making while liquidity demanders tend to pay a larger transaction cost because of it. From a policy perspective, then, it would be the first question whether or not this wealth transfer is rationalizable. If tick size increase does not improve liquidity as much as it adds up to transaction costs, the policy idea of building up liquidity for small-cap stocks by means of a tick size increase would lose its ground.

A straightforward approach to this policy question would be weighing differential effects of tick size increase on quoted spread and depth. An increase in quoted spread reflects an increase in transaction costs liquidity demanders are required to pay for an improved liquidity environment. On the other hand, an increase in depth is the output of an improved liquidity environment benefiting liquidity demanders. If the effect of the latter outweighs that of the former, then this policy idea is, at least, rationalizable. It is also important to examine whether there is an unintended side effect of tick size increase on the market. If, for example, the market becomes more volatile due to the tick size increase, then this should be also taken into account.

This paper carries all the three outcomes in the form of percentage quoted spread, consolidated displayed depth, and high-low volatility. Importantly, the distributions of the ITE test statistics, discussed in the previous section, enable to weigh differing policy effects of the tick size increase on percentage quoted spread and consolidated displayed depth in a statistical sense. The testing results for Treatment Group 1 in Table 5 in particular show that more than 83% of the pilot stocks receive significant increases in consolidated displayed depth due to the tick size increase. On the other hand, only less than half of them reveal significant increase in percentage quoted spread. Those collectively imply that the tick size increase does not necessarily lead to increases in percentage quoted spread but improvement in consolidated displayed depth is widely observed. Note that consolidated displayed depth used in this paper only counts displayed depth at NBBO. In turn, the estimated effects on it may convey the lower bound of the total improvement in depth due to the tick size increase. Counting all the hidden and/or near-NBBO depth, the total improvement in depth can be greater. Equally important, the results on high-low volatil-

ity, which measures a short-run volatility, do not present a significant disturbance in link to the tick size increase. In sum, policy learning of this paper concludes that increasing tick size from penny to nickel can be an effective way of improving liquidity for small-cap stocks.

7. Concluding Remarks

In this paper, I investigate TSPP, the latest RCT conducted in the U.S stock market. Using a ML approach, I estimate ITE for pilot stocks and test its significance at the individual stock level to unravel policy effects beyond ATE. The results of the ML approach show that the tick size increase from one penny to one nickel under TSPP impacts on percentage quoted spread unevenly across affected pilot stocks. Only less than half of the pilot stocks show significance. On the other hand, the effects of the tick size change is comprehensively positive on consolidated displayed depth across affected pilot stocks. Meanwhile, the tick size change do not significantly affect short-term volatility for almost every pilot stock, measured by high-low volatility on half-hour intervals. Furthermore, the individual results reveal no significant spillover effect of policy interventions over the control group.

I also look into individual-specific characteristics to understand differing effects of the tick size change in the cross-section. Consistent with previous findings in the literature, tick constrainedness in the pre-intervention periods accounts for policy effect heterogeneity on percentage quoted spread. However, tick constrainedness lacks of explanatory power for consolidated displayed depth. Percentage realized spreads in the pre-intervention periods in contrast explain heterogeneous policy effects on the both outcomes in the cross-section, revealing that the lower percentage realized spread in the pre-intervention periods, the more likely significantly affected by the tick size change in the post-intervention periods.

In addition, I estimate panel-data ATEs from a ML approach that does not involve the control group of TSPP at any stage. It shows that ATE estimates drawn from this ML approach are statistically indistinguishable from those estimated from popular panel-data regression models that exploit the RCT design. This result implies that a big-data- and ML-based predictive model can be as good as a RCT approach for causal inference on ATE, which has long been believed to be the gold standard for casual inference.

References

Abadie, Alberto. (2019). "Using synthetic controls: Feasibility, data requirements, and methodological aspects." *Journal of Economic Literature*. forthcoming.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. (2010). "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program." *Journal of the American Statistical Association* 105(490): 493–505.

——— (2015). "Comparative politics and the synthetic control method." *American Journal of Political Science* 59(2): 495–510.

Ahn, Hee-Joon, Charles Q Cao, and Hyuk Choe. (1996). "Tick size, spread, and volume." *Journal of Financial Intermediation* 5(1): 2–22.

——— (1998). "Decimalization and competition among stock markets: Evidence from the Toronto Stock Exchange cross-listed securities." *Journal of Financial Markets* 1(1): 51–87.

Albuquerque, Rui A, Shiyun Song, and Chen Yao. (2019). "The price effects of liquidity shocks: A study of SEC's tick-size experiment." Unpublished Working Paper. https://ssrn.com/abstract=3081125.

Angel, James J. (1997). "Tick size, share prices, and stock splits." *Journal of Finance* 52(2): 655–681.

Anshuman, V Ravi and Avner Kalay. (1998). "Market making with discrete prices." *Review of Financial Studies* 11(1): 81–109.

Athey, Susan. (2017). "Beyond prediction: Using big data for policy problems." *Science* 355(6324): 483–485.

Athey, Susan and Guido W Imbens. (2017). "The econometrics of randomized experiments." *Handbook of Economic Field Experiments* 1: 73–140.

——— (2019). "Machine learning methods that economists should know about." *Annual Review of Economics* 11: 685–725.

Bacidore, Jeffrey. (1997). "The impact of decimalization on market quality: An empirical investigation of the Toronto Stock Exchange." *Journal of Financial Intermediation* 6(2): 92–120.

Bacidore, Jeffrey, Robert H Battalio, and Robert H Jennings. (2003). "Order submission strategies, liquidity supply, and trading in pennies on the New York Stock Exchange." *Journal of Financial Markets* 6(3): 337–362.

Banerjee, Abhijit Vinayak, Esther Duflo, and Michael Kremer. (2016). "The influence of randomized controlled trials on development economics research and on development policy." *The State of Economics, The State of the World*.

Bartlett, Robert P and Justin McCrary. (2017). "Subsidizing liquidity with wider ticks: Evidence from the tick size pilot study." Unpublished Working Paper. https://ssrn.com/abstract=3076257.

Bessembinder, Hendrik. (2003). "Trade execution costs and market quality after decimalization." *Journal of Financial and Quantitative Analysis* 38(4): 747–777.

Bessembinder, Hendrik, Jia Hao, and Kuncheng Zheng. (2015). "Market making contracts, firm value, and the IPO decision." *Journal of Finance* 70(5): 1997–2028.

Bloomfield, Robert, Maureen O'hara, and Gideon Saar. (2005). "The 'make or take' decision in an electronic market: Evidence on the evolution of liquidity." *Journal of Financial Economics* 75(1): 165–199.

Boehmer, Ekkehart, Charles M Jones, and Xiaoyan Zhang. (2020). "Potential pilot problems: Treatment spillovers in financial regulatory experiments." *Journal of Financial Economics* 135(1): 68–87.

Bollen, Nicolas PB and Jeffrey A Busse. (2006). "Tick size and institutional trading costs: Evidence from mutual funds." *Journal of Financial and Quantitative Analysis* 41(4): 915–937.

Brogaard, Jonathan and Jing Pan. (2019). "Dark Trading and the Fundamental Information in Stock Prices." Unpublished Working Paper. https://ssrn.com/abstract=3281472.

Burlig, Fiona, Christopher Knittel, David Rapson, Mar Reguant, and Catherine Wolfram. (2017). "Machine learning from schools about energy efficiency." Unpublished Working Paper. https://www.nber.org/papers/w23908.

Buti, Sabrina, Barbara Rindi, and Ingrid M Werner. (2011). "Diving into dark pools." Unpublished Working Paper. https://ssrn.com/abstract=1630499.

Chakravarty, Sugato, Venkatesh Panchapagesan, and Robert A Wood. (2005). "Did decimalization hurt institutional investors?" *Journal of Financial Markets* 8(4): 400–420.

Chakravarty, Sugato, Robert A Wood, and Robert A Van Ness. (2004). "Decimals and liquidity: A study of the NYSE." *Journal of Financial Research* 27(1): 75–94.

Chernozhukov, Victor, Kaspar Wuthrich, and Yinchu Zhu. (2020). "Practical and robust t-test based inference for synthetic control and related methods." Unpublished Working Paper. https://arxiv.org/abs/1812.10820.

Chordia, Tarun, Richard Roll, and Avanidhar Subrahmanyam. (2008). "Liquidity and market efficiency." *Journal of Financial Economics* 87(2): 249–268.

Chung, Kee H and Chairat Chuwonganant. (2002). "Tick size and quote revisions on the NYSE." *Journal of Financial Markets* 5(4): 391–410.

Chung, Kee H, Albert J Lee, and Dominik Rösch. (2020). "Tick size, liquidity for small and large orders, and price informativeness: Evidence from the Tick Size Pilot Program." *Journal of Financial Economics* 136(3): 879–899.

Chung, Kee H, Bonnie F Van Ness, and Robert A Van Ness. (1999). "Limit orders and the bid–ask spread." *Journal of Financial Economics* 53(2): 255–287.

Comerton-Forde, Carole, Vincent Grégoire, and Zhuo Zhong. (2019). "Inverted fee structures, tick size, and market quality." *Journal of Financial Economics* 134(1): 141–164.

Cox, Justin, Bonnie Van Ness, and Robert Van Ness. (2019). "Increasing the tick: Examining the impact of the tick size change on maker-taker and taker-maker market Models." *Financial Review* 54(3): 417–449.

Currie, Janet, Henrik Kleven, and Esmée Zwiers. (2020). "Technology and big data are changing economics: mining text to track methods." Unpublished Working Paper. https://www.nber.org/papers/w26715.

Deaton, Angus. (2010). "Instruments, randomization, and learning about development." *Journal of economic literature* 48(2): 424–55.

——— (2019). "Randomization in the tropics revisited: a theme and eleven variations." Unpublished Working Paper. https://scholar.princeton.edu/deaton/publications/randomization-tropics-revisited-theme-and-eleven-variations-randomized.

Deaton, Angus and Nancy Cartwright. (2018). "Understanding and misunderstanding randomized controlled trials." *Social Science & Medicine* 210: 2–21.

Doudchenko, Nikolay and Guido W Imbens. (2016). "Balancing, regression, difference-in-differences and synthetic control methods: A synthesis." Unpublished Working Paper. https://www.nber.org/papers/w22791.

Farley, Ryan, Eric K Kelley, and Andy Puckett. (2018). "Dark trading volume and market quality: A natural experiment." Unpublished Working Paper. https://ssrn.com/abstract=3088715.

Foster, F Douglas and Subramanian Viswanathan. (1993). "Variations in trading volume, return volatility, and trading costs: Evidence on recent price formation models." *Journal of Finance* 48(1): 187–211.

Foucault, Thierry, Ohad Kadan, and Eugene Kandel. (2005). "Limit order book as a market for liquidity." *Review of Financial Studies* 18(4): 1171–1217.

Freedman, David A. (2006). "Statistical models for causation: what inferential leverage do they provide?" *Evaluation review* 30(6): 691–713.

Goettler, Ronald L, Christine A Parlour, and Uday Rajan. (2005). "Equilibrium in a dynamic limit order market." *Journal of Finance* 60(5): 2149–2192.

Goldstein, Michael A and Kenneth A Kavajecz. (2000). "Eighths, sixteenths, and market depth: changes in tick size and liquidity provision on the NYSE." *Journal of Financial Economics* 56(1): 125–149.

Griffith, Todd G and Brian S Roseman. (2019). "Making cents of tick sizes: The effect of the 2016 US SEC tick size pilot on limit order book liquidity." *Journal of Banking & Finance* 101: 104–121.

Griffiths, Mark D, Brian F Smith, D Alasdair S Turnbull, and Robert W White. (2000). "The costs and determinants of order aggressiveness." *Journal of Financial Economics* 56(1): 65–88.

Hansen, Peter, Yifan Li, Asger Lunde, and Andrew Patton. (2017). "Mind the Gap: An Early Empirical Analysis of SEC's Tick Size Pilot Program." Unpublished Working Paper.

Harris, Lawrence. (1994). "Minimum price variations, discrete bid–ask spreads, and quotation sizes." *Review of Financial Studies* 7(1): 149–178.

——— (1996). "Does a large minimum price variation encourage order exposure?". Unpublished Working Paper. https://pdfs.semanticscholar.org/c026/5eb867f8f89cbc81670cdf63cc7afc66f4a1.pdf.

——— (1997). "Decimalization: A review of the arguments and evidence." Unpublished Working Paper. https://pdfs.semanticscholar.org/7e55/279415ae4d11f701ff93bbe5585205faeb48.pdf.

——— (1999). "Trading in pennies: a survey of the issues." Unpublished Working Paper. https://pdfs.semanticscholar.org/06cc/896a7b96002abcee5b3cea5148570707b08d.pdf.

Harris, Lawrence and Joel Hasbrouck. (1996). "Market vs. limit orders: The SuperDOT evidence on order submission strategy." *Journal of Financial and Quantitative analysis* 31(2): 213–231.

Heckman, James J. (2008). "Econometric causality." *International statistical review* 76(1): 1–27.

——— (2020). "Randomization and Social Policy Evaluation Revisited." Unpublished Working Paper. https://www.nber.org/papers/t0107.

Holden, Craig W and Stacey Jacobsen. (2014). "Liquidity measurement problems in fast, competitive markets: Expensive and cheap solutions." *Journal of Finance* 69(4): 1747–1785.

Holland, Paul W. (1986). "Statistics and causal inference." *Journal of the American Statistical Association* 81(396): 945–960.

Hollifield, Burton, Robert A Miller, and Patrik Sandås. (2004). "Empirical analysis of limit order markets." *The Review of Economic Studies* 71(4): 1027–1063.

Hu, Edwin, Paul Hughes, John Ritter, Patti Vegella, and Hao Zhang. (2018). "Tick size pilot plan and market quality." Unpublished Working Paper. https://www.sec.gov/dera/staff-papers/white-papers/dera_wp_tick_size-maet_quality.

Imbens, Guido W and Donald B Rubin. (2015) *Causal inference in statistics, social, and biomedical sciences*: Cambridge University Press.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. (2015). "Prediction policy problems." *American Economic Review* 105(5): 491–95.

Kwan, Amy, Ronald Masulis, and Thomas H McInish. (2015). "Trading rules, competition for order flow and market fragmentation." *Journal of Financial Economics* 115(2): 330–348.

Kye, Hyungil and Bruce Mizrach. (2019). "The Off-Exchange Routing Decision." Unpublished Working Paper. https://ssrn.com/abstract=3440911.

Lee, Charles MC and Mark J Ready. (1991). "Inferring trade direction from intraday data." *Journal of Finance* 46(2): 733–746.

Lee, Charles MC and Edward M Watts. (2018). "Tick Size Tolls: Can a Trading Slowdown Improve Earnings News Discovery?". Unpublished Working Paper. https://ssrn.com/abstract=3263778.

Li, Xiongshi, Mao Ye, and Miles Zheng. (2019). "Market Structure and Corporate Payout Policy: Evidence from a Controlled Experiment." Unpublished Working Paper. https://ssrn.com/abstract=3254585.

Lin, Ji-Chai, Gary C Sanger, and G Geoffrey Booth. (1995). "Trade size and components of the bid-ask spread." *Review of Financial Studies* 8(4): 1153–1183.

Lin, Yiping, Peter L Swan, and Vito Mollica. (2018). "Reg NMS and Minimum-Tick Distort the Market in Opposing Directions: Theory and Market Experimental Evidence." Unpublished Working Paper. https://ssrn.com/abstract=2913555.

Madhavan, Ananth, Matthew Richardson, and Mark Roomans. (1997). "Why do security prices change? A transaction-level analysis of NYSE stocks." *Review of Financial Studies* 10(4): 1035–1064.

McDonald, John H. (2014) *Handbook of Biological Statistics*: Sparky House Publishing, Baltimore, Maryland, 3rd edition.

McInish, Thomas H and Robert A Wood. (1992). "An analysis of intraday patterns in bid/ask spreads for NYSE stocks." *Journal of Finance* 47(2): 753–764.

Mullainathan, Sendhil and Jann Spiess. (2017). "Machine learning: an applied econometric approach." *Journal of Economic Perspectives* 31(2): 87–106.

Penalva, José and Mikel Tapia. (2017). "Revisiting Tick Size: Implications from the SEC Tick Size Pilot." Unpublished Working Paper. https://ssrn.com/abstract=2994892.

Ranaldo, Angelo. (2004). "Order aggressiveness in limit order book markets." *Journal of Financial Markets* 7(1): 53–74.

Ready, Mark. (2014). "Determinants of volume in dark pool crossing networks." Unpublished Working Paper. https://ssrn.com/abstract=1361234.

Rindi, Barbara and Ingrid M Werner. (2019). "US tick size pilot." Unpublished Working Paper. https://ssrn.com/abstract=3041644.

Ronen, Tavy and Daniel G Weaver. (2001). "'Teenies' anyone?" *Journal of Financial Markets* 4(3): 231–260.

SEC, U.S. Securities and Exchange Commission. (2012) "Report to Congress on Decimalization." As Required by Section 106 of the Jumpstart Our Business Startups Act. https://www.sec.gov/news/studies/2012/decimalization-072012.pdf.

Seppi, Duane J. (1997). "Liquidity provision with limit orders and a strategic specialist." *Review of Financial Studies* 10(1): 103–150.

Stoll, Hans R and Robert E Whaley. (1990). "Stock market structure and volatility." *Review of Financial Studies* 3(1): 37–71.

Thomas, Jacob K, Frank Zhang, and Wei Zhu. (2018). "Off-exchange trading and post earnings announcement drift." Unpublished Working Paper. https://ssrn.com/abstract=3223528.

Van Ness, Bonnie F, Robert A Van Ness, and Stephen W Pruitt. (2000). "The impact of the reduction in tick increments in major US markets on spreads, depth, and volatility." *Review of Quantitative Finance and Accounting* 15(2): 153–167.

Varian, Hal R. (2014). "Big data: New tricks for econometrics." *Journal of Economic Perspectives* 28(2): 3–28.

Ye, Mao, Miles Zheng, and Wei Zhu. (2019). "Price Discreteness and Investment to Price Sensitivity." Unpublished Working Paper. https://ssrn.com/abstract=3517305.

Young, Alwyn. (2019). "Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results." *Quarterly Journal of Economics* 134(2): 557–598.

A. Tables

Table 1: Prior Empirical Works on the Tick Size Pilot Program

This table lists prior empirical papers that take the Tick Size Pilot Program as the main or partial theme in their works in the form of statistical analyses. The first column represents papers by authors along with year lastly updated or published (*). The second column shows outcome variables utilized by the papers. Definition of variables can be specific paper-by-paper. A typical set of variables are the following: a) market quality measures: (nominal or percentage) quoted spread, realized spread, effective spread, price impacts, displayed depth, or the like; b) volatility measures: midquote realized volatility, percentage high-low midquote range; c) market efficiency: proxied by variance ratios or autocorrelation of short-term returns; d) trading activities: aggregate (share or dollar) trading volume, aggregate trades, trade size, etc.; e) algorithmic trading proxies: some combinations of counting measures from trading and quoting messages; f) market shares of trading venues: (share or dollar) trading volume of a certain venue(s) divided by consolidated trading volume. The third column documents papers' sample periods corresponding to the periods of statistical analyses. For instance, one-year window represents the sample period covering one year before and after the policy intervention. The last column shows empirical frameworks of papers, where a standard two-way fixed-effect model is denoted by difference-in-difference as they essentially represent the identical parameters of interest.

Papers	Outcome Variables	Sample Period	Empirical Framework
Albuquerque, Song, and Yao (2019)	abnormal returns, market quality measures	Apr. 2016 - Apr. 2017	difference-in-difference
Bartlett and McCrary (2017)	market quality measures	Mar. 2016 - Jun. 2017	group mean comparisons
Brogaard and Pan (2019)	trading volume market shares	120-day window	difference-in-difference
Chung, Lee, and Rösch (2020)*	market quality, volatility, market efficiency, and trading activity measures	one-year window	difference-in-difference
Comerton-Forde, Grégoire, and Zhong (2019)*	trading volume market shares	Sept. 1 - Dec. 2016	difference-in-difference
Cox, Van Ness, and Van Ness (2019)*	algorithmic trading proxies, trading activity measures	30-day window	difference-in-difference
Farley, Kelley, and Puckett (2018)	market quality, volatility, and market efficiency measures	20-day window	difference-in-difference
Griffith and Roseman (2019)*	market quality measures	six-week window	difference-in-difference
Hansen et al. (2017)	market quality, volatility, and trading activity measures	three-month window	difference-in-difference
Hu et al. (2018)	market quality, volatility, market efficiency, and trading activity measures	four-month window	difference-in-difference
Lee and Watts (2018)	algorithmic trading proxies, market quality and trading activity measures, etc.	two-year window	difference-in-difference
Li, Ye, and Zheng (2019)	corporate payout measures	eight-quarter window	difference-in-difference
Lin, Swan, and Mollica (2018)	market quality, volatility, market efficiency, and trading activity measures	two-month window	difference-in-difference
Penalva and Tapia (2017)	market quality, volatility, and trading activity measures	Sept. 6 - Nov. 30, 2016	difference-in-difference
Rindi and Werner (2019)	market quality, volatility, and trading activity measures	eight-week window	difference-in-difference
Thomas, Zhang, and Zhu (2018)	returns	eight-quarter window	difference-in-difference
Ye, Zheng, and Zhu (2019)	Tobin's q	two-year window	difference-in-difference

Table 2: Sample Construction of Pilot Stocks

This table shows the sampling process for pilot stocks based on the official symbol list published on the FINRA's Tick Size Pilot Program website. The first column states stock deletion rules with related data sets in brackets; the second through fourth columns represent the number of pilot stocks and their relative changes in parentheses deducted by the deletion rules; the last column provides supplementary explanations regarding the deletion rules.

The data sets used in this process are the FINRA's pilot symbols list data (**F1**, **F2**, and **F3**), daily **CRSP**, and the half-hour inside quote (**HHIQ**), which is based on the NBBO-sorted TAQ quote data with the HJ algorithm in place. **F1** contains the official list of the pilot stocks published by FINRA on Oct. 28, 2016; **F2** traces changes of the pilot stocks related to institutional status and experimental designs over the experimental periods; **F3** is similar to **F2** but tracks the changes over the pre-pilot periods in Mar. 8 - Sept. 2, 2016.

Deletion Rules [Data Sets]	Control Group	Treatment Group 1	Treatment Group 2	Treatment Group 3	Descriptions
Initial assignment in the experiment [**F1**]	1,196	397	395	395	12 test ticker symbols excluded.[a]
Changes in ticker symbols [**F2**, **F3**]	1,155 (41)	381 (16)	374 (21)	376 (19)	Change code: `C` or `S`.
Switching listing exchanges [**F2**, **F3**]	1,140 (15)	376 (5)	372 (2)	370 (6)	Change code: `T` or `U`.
Falling to control group [**F2**]	1,140 (0)	361 (15)	359 (13)	350 (20)	Change code: `P`
Early dropout [**F2**]	1,135 (5)	360 (1)	358 (1)	350 (0)	Deleted prior to Nov. 1, 2016.
Erroneous or incomplete records [**F2**]	1,130 (5)	357 (3)	358 (0)	349 (1)	Eight symbols not fully tractable.[b]
Minimum survival periods [**F2**]	1,038 (92)	335 (22)	333 (25)	323 (26)	Surviving through Oct. 2017 or longer.
Daily trading volume [**CRSP**]					5,000-share daily volume for more than two-thirds of the trading days in each sample.
- Training sample	852 (186)	284 (51)	268 (65)	258 (65)	
- Post-intervention sample	842 (10)	282 (2)	266 (2)	255 (3)	
- Pre-intervention sample	838 (4)	281 (1)	264 (2)	254 (1)	
Late starters in the training sample [**CRSP**]	815 (23)	273 (1)	259 (1)	245 (1)	Having the 1st valid training day later than Jan. 16.
Poorly matched stocks in data sets [**CRSP**, **HHIQ**]	814 (1)	272 (1)	259 (0)	244 (1)	Three symbols.[c]
No valid NBBO quote within one hour [**HHIQ**]	811 (3)	272 (0)	259 (0)	243 (1)	Four symbols.[d]
Liquidity events (e.g., M&A, public offerings)	810 (1)	272 (0)	257 (2)	242 (1)	Four symbols.[e]

[a] `ATEST, ATEST A, ATEST B, ATEST C, NTEST, NTEST A, NTEST B, NTEST C, ZAZZT, ZBZZT, ZCZZT, ZVZZT`.
[b] `AST, FOR, KLDX, NSU, SNOW, TIK, XOXO, XTLY`.
[c] `AAMC, ATRI, NWLI`.
[d] `DHIL, GHC, MLAB, TPL`.
[e] `FGL, FIG, RATE, SRNE`.

F1: https://www.finra.org/sites/default/files/Tick_Pilot_Test_Group_Assignments.txt.
F2: http://tsp.finra.org/finra_org/ticksizepilot/TSPilotChanges.txt.
F3: https://www.finra.org/sites/default/files/TSPrePilotChanges_20160902.txt.

Table 3: Sample Construction of Donor-Pool Stocks

This table shows the sampling process for donor pools. The first column states sampling rules; the second column counts the number of the stocks and their relative changes in parentheses caused by the sampling rules; the third column represents relative daily trading volume shares following the sampling process; the last column writes supplementary explanations on the sampling rules.

Sampling Rules	Securities (change)	Volume Share (%)	Descriptions
Using **CRSP** data:			
U.S.-listed stocks	8,451	-	Stocks uniquely identified with `TSYMBOL` and listed in `EXCD` $= 1, 2, 3,$ or 4 over the whole sample period of Jan. 2015 - Jul. 2017.
Non-pilot stocks	5,994 (-2,457)	100.00	Exclude the stocks whose ticker symbols are, partly or fully, found in the ticker symbol list of the pilot stocks.
5,000-share trading volume	5,955 (-39)	99.99	Apply 5,000-share cut-off to stocks-days.
Abnormality filters:			
- No spilt events	5,480 (-475)	87.60	Exclude the stocks that have a day-to-day change in either `CFACPR` or `CFACSHR`, with exception of the case of less than 10% overnight returns.
- No excessive overnight returns	5,006 (-474)	81.52	Exclude the stocks that have a more than 30% overnight return.
- No abrupt change in outstanding shares	4,552 (-454)	78.44	Exclude the stocks that have a day-to-day change of more than 50% in `SHROUT`.
Comparable trading-day sequence	2,331 (-2,221)	72.76	Include only the stocks that have the same trading-day sequence over the whole sample period as at least one pilot stock sampled from Table 2.
Positive volatility	2,307 (-24)	72.58	Exclude the stocks that have standard deviation of the daily returns over the whole sample period less than 15 basis points.
Winsorization	2,220 (-88)	64.48	Based on opening price and market capitalization on Sept. 1, 2016, or the closest last trading day, include only the stocks whose values of the both variables lying between the 1% and 99% quantiles.
Using **Daily TAQ** quote data, NBBO-sorted by the HJ algorithm:			
The first valid NBBO within one hour from the opening bell	2,220	-	Delete stocks-days that have the first valid NBBO quote later than 10:00 AM.
Stationary outcomes*			
Percentage quoted spread	1,280	-	Include only the stocks that show no sign of being influenced by the pilot program on time series of outcomes over pre- and post-intervention periods.
Consolidated displayed depth	1,343	-	
High-low volatility	2,045	-	

* Stationarity of outcomes is gauged through three tests. The first test is whether outcomes follow a random-walk over the pre-intervention period; the second is a standard two-sample t-test for the long-run means of outcomes between the pre- and post-intervention periods; the last one is a structural break test on outcomes fitted to an autoregressive model between the pre- and post-intervention periods. The stocks showing statistical evidence of following a random-walk process or rejecting the both nulls of the long-run means and structural breaks tests are excluded from the donor pool for a given outcome.

Table 4: Descriptive Statistics

This table shows descriptive statistics of the three outcomes of interest, percentage quoted spread, consolidated displayed depth, and high-low volatility. The sample means and sample standard deviations in parentheses are computed over the three sample periods, training sample (Jan. - Dec. 2015), pre-intervention sample (Jan. - Sept. 2016), and post-intervention sample (Nov. 2016 - Jul. 2017). The pilot stocks in either treatment or control group(s) are sampled according to the rules in Table 2, and the donor pools are constructed following the procedure in Table 3.

Outcomes	**Control Group**	**Treatment Group 1**	**Treatment Group 2**	**Treatment Group 3**	**Donor Pools**
Percentage Quoted Spread (bps)					
Training Period	52.69	48.35	47.33	48.38	15.31
	(78.65)	(67.97)	(67.39)	(69.05)	(22.20)
Pre-Intervention Period	54.39	51.29	49.87	51.09	16.34
	(83.88)	(75.16)	(73.87)	(78.69)	(23.77)
Post-Intervention Period	50.23	64.44	59.84	61.44	14.06
	(75.95)	(70.14)	(65.65)	(66.17)	(20.58)
Consolidated Displayed Depth (shares)					
Training Period	636.93	754.75	603.82	738.49	3781.00
	(1528.05)	(2273.73)	(1125.12)	(2082.36)	(18567.31)
Pre-Intervention Period	629.45	681.78	627.12	683.55	4571.91
	(1124.56)	(1491.11)	(1093.54)	(1375.77)	(21440.13)
Post-Intervention Period	784.86	2990.71	2860.76	3893.55	5066.17
	(1931.35)	(6825.19)	(5722.33)	(10848.51)	(29158.74)
High-Low Volatility (bps)					
Training Period	82.07	79.57	79.19	76.50	45.82
	(82.46)	(77.46)	(76.47)	(75.27)	(54.22)
Pre-Intervention Period	86.79	83.02	84.02	81.50	48.97
	(89.80)	(82.43)	(84.09)	(81.32)	(57.20)
Post-Intervention Period	76.17	62.95	64.76	64.34	37.86
	(76.60)	(69.00)	(71.04)	(70.27)	(43.34)

Table 5: Inference on Individual Treatment Effects

This shows results of statistical inference based on the ITE test statistics (9). For each pilot stock i, ITE is separately tested one-by-one. The p-value of each testing, P, is computed against the null distributions, drawn in Figure 4. Also, in consideration of the multiple testing problem, testing results based on the Benjamini-Hochberg procedure with false discovery rate $\alpha = 0.05$ within each group are reported in P_{BH}.

For the three outcomes of interest, the table counts the number of the pilot stocks within each group that show statistical significance at 5% level, i.e., p-values smaller than 0.05, along with percentage of such stocks in parentheses.

	N	$P < .05$	$P_{BH} < .05$
Perecentage Quoted Spread			
Treatment Group 1	272	144 (52.94%)	131 (48.16%)
Treatment Group 2	257	124 (48.25%)	112 (43.58%)
Treatment Group 3	242	124 (51.24%)	111 (45.87%)
Control Group	810	113 (13.93%)	27 (3.33%)
Consolidated Displayed Depth			
Treatment Group 1	272	230 (84.56%)	228 (83.82%)
Treatment Group 2	257	224 (87.16%)	223 (86.77%)
Treatment Group 3	242	215 (88.84%)	212 (87.60%)
Control Group	810	81 (10.00%)	18 (2.22%)
High-Low Volatility			
Treatment Group 1	272	36 (13.24%)	10 (3.68%)
Treatment Group 2	257	29 (11.28%)	10 (3.89%)
Treatment Group 3	242	20 (8.26%)	4 (1.65%)
Control Group	810	64 (7.90%)	15 (1.85%)

Table 6: Panel Data Regressions

This reports panel-data regression results for percentage quoted spread, consolidated displayed depth, and high-low volatility with half-hour sample data over the pre-intervention period (9 months; Jan. - Sept. 2016) and the post-intervention period (9 months; Nov. 2016 - Jul. 2017). $G1_i$, $G2_i$, and, $G3_i$ are the treatment status indicators for a pilot stock on the three treatment groups – if stock i belongs to one of the treatment groups, then the corresponding treatment group takes one and zero otherwise; $Pilot_t$ is the treatment period indicator that takes one if half-hour time index t is in the post-intervention period and zero otherwise; VIX_t is the Chicago Board Options Exchange's Volatility Index at opening of each half-hour interval t; Capitalization_i is the market value for stock i, which is time-invariant, calculated as opening price times the number of outstanding shares based on Jan. 4, 2016, the first trading day in the regression data.

(M1) represents fixed effect model (12), and (M2), (M3), and (M3) are difference-in-difference models (13) with different choice of covariates: no covariate, inclusion only of VIX, and inclusion both of VIX and log capitalization. Standard errors are adjusted with clustering in stocks and dates.

	Percentage Quoted Spread				**Consolidated Displayed Depth**				**High-Low Volatility**			
	(M1)	(M2)	(M3)	(M4)	(M1)	(M2)	(M3)	(M4)	(M1)	(M2)	(M3)	(M4)
$G1_i \cdot Pilot_t$	17.77***	17.31***	17.32***	17.54***	2154.15***	2153.08***	2153.05***	2154.47***	-9.38***	-9.44***	-9.43***	-9.37***
	(2.26)	(2.19)	(2.19)	(2.24)	(291.28)	(290.42)	(290.42)	(290.53)	(1.61)	(1.59)	(1.59)	(1.60)
$G2_i \cdot Pilot_t$	14.69***	14.17***	14.17***	14.40***	2075.61***	2077.52***	2077.48***	2078.94***	-8.55***	-8.63***	-8.61***	-8.55***
	(2.37)	(2.30)	(2.30)	(2.33)	(247.22)	(246.63)	(246.63)	(246.60)	(1.51)	(1.49)	(1.41)	(1.49)
$G3_i \cdot Pilot_t$	14.68***	14.54***	14.54***	14.52***	3052.89***	3052.68***	3052.66***	3052.50***	-6.41***	-6.52***	-6.51***	-6.52***
	(2.51)	(2.39)	(2.39)	(2.45)	(537.77)	(535.54)	(535.54)	(535.55)	(1.42)	(1.41)	(1.41)	(1.41)
$Pilot_t$		-4.14***	1.51	1.18		155.03***	104.62***	102.51***		-10.59***	7.89***	7.81***
		(1.09)	(0.96)	(0.99)		(24.97)	(24.91)	(24.94)		(1.83)	(1.52)	(1.51)
VIX_t			1.24***	1.23***			-11.05***	-11.11***			4.05***	4.05***
			(0.10)	(0.10)			(1.84)	(1.84)			(0.23)	(0.23)
$\log(MktCap)_i$				-33.86***				-218.59***				-8.91***
				(1.01)				(49.31)				(0.65)
Unit FE	Yes	No	No	No	Yes	No	No	No	Yes	No	No	No
Time FE	Yes	No	No	No	Yes	No	No	No	Yes	No	No	No
Group FE	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
$\bar{R}^2$	0.4471	0.0039	0.0068	0.2425	0.4139	0.0668	0.0669	0.0700	0.1704	0.0114	0.0389	0.0535

standard errors in parenthesis clustered by stocks and dates
* p<0.05, ** p<0.01, *** p<0.001; Observations: 7,622,914

Table 7: Average Treatment Effects: ML vs. RCT

This shows estimates of panel-data ATEs for the three treatment groups and their standard errors in parentheses on the three outcome variables, percentage quoted spread, consolidated displayed depth, and high-low volatility, obtained from machine learning (ML) and panel-data regression exploiting the RCT design (RCT). The ML estimator of ATE and its standard errors are defined in (10) and (11). The RCT estimates and their standard errors are drawn from (M1) in Table 6, a two-way fixed effect panel data model with inclusion of the treatment indicators only. As reference, t-values for difference of ATEs between ML and REG are reported, computed as difference of the ATE estimates divided by the sum of their respective standard errors.

	Percentage Quoted Spread			**Consolidated Displayed Depth**			**High-Low Volatility**		
	ML	RCT	Diff [t-val]	ML	RCT	Diff [t-val]	ML	RCT	Diff [t-val]
$\widehat{ATE}_1$	18.64	17.77	0.87	2184.52	2154.15	30.37	-6.83	-9.38	2.55
	(2.10)	(2.26)	[0.20]	(283.09)	(291.28)	[0.05]	(1.21)	(1.61)	[0.90]
$\widehat{ATE}_2$	14.95	14.69	0.26	2130.10	2075.61	54.49	-6.09	-8.55	2.46
	(2.12)	(2.37)	[0.06]	(239.22)	(247.22)	[0.11]	(1.09)	(1.51)	[0.95]
$\widehat{ATE}_3$	15.12	14.68	0.44	3113.76	3052.89	60.87	-4.11	-6.41	2.3
	(2.40)	(2.51)	[0.09]	(520.25)	(537.77)	[0.06]	(0.96)	(1.42)	[0.97]

Table 8: Policy Effects Heterogeneity

This summarizes results of analysis of policy effect heterogeneity. It is based on statistical significance of individual treatment effects for pilot stocks in the treatment groups. $S_i \in \{0,1\}$ takes one if pilot stock i in the treatment groups shows statistical significance at the 5% level and zero otherwise. S_i is Probit-regressed on individual characteristics X_i. Then, an estimate of partial effect with respect to k-th characteristic at sample means and its standard error are reported below.

To be specific, a Probit model is given by:

$$S_i = \mathbf{1}\{X_i'\theta > u_i\}, \quad u_i \sim \mathcal{N}(0,1)$$

The estimator of the partial effect for continuous variable X_k is defined as:

$$\widehat{PE}_k \equiv \phi(\bar{X}_i'\hat{\theta})\hat{\theta}_k, \quad \phi(\cdot) : \text{density function of } u_i$$

The estimator of the partial effect for binary variable X_k is defined as:

$$\widehat{PE}_k \equiv \Phi(\bar{X}_i'\hat{\theta})\big|_{X_k=1} - \Phi(\bar{X}_i'\hat{\theta})\big|_{X_k=0}, \quad \Phi(\cdot) : \text{ distribution function of } u_i$$

Individual characteristics X_i includes the tick constrainedness indicator that takes one if average daily time-weighted quoted spread for stock i in the pre-intervention period (Jan. - Sept. 2016) is less than \$0.05 and zero otherwise, and several other time-series means as individual-specific characteristics. For each stock i time-series means are computed over the pre-intervention period on percentage realized spread measured at various time horizons (30-seconds, one-minutes, and five-minutes), daily market capitalization (MktCap), daily opening price, and daily trading volume.

	Percentage Quoted Spread			**Consolidated Displayed Depth**		
Tick Constrained	0.5208***	0.5309***	0.5507***	-0.0413	-0.0403	-0.0385
	(0.0486)	(0.0475)	(0.0457)	(0.0322)	(0.0321)	(0.0317)
Percentage Realized Spread (30-sec; bps)	-0.0123***			-0.0024**		
	(0.0030)			(0.0008)		
Percentage Realized Spread (1-min; bps)		-0.0115***			-0.0025**	
		(0.0031)			(0.0008)	
Percentage Realized Spread (5-min; bps)			-0.0095***			-0.0027**
			(0.0031)			(0.0009)
log(MktCap)	0.0071	0.0138	0.0258	0.0167	0.0181	0.0210
	(0.0402)	(0.0401)	(0.0388)	(0.0175)	(0.0174)	(0.0173)
log(Volume)	0.0722	0.0804*	0.1040*	0.0634***	0.0637***	0.0648***
	(0.0371)	(0.0368)	(0.0355)	(0.0153)	(0.0153)	(0.0151)
1/Price	1.8643**	1.7371**	1.3641**	0.4656*	0.4790*	0.4983*
	(0.5993)	(0.5924)	(0.5648)	(0.2233)	(0.2265)	(0.2282)
Group 2	-0.0540	-0.0533	-0.0538	0.0202	0.0203	0.0197
	(0.0532)	(0.0533)	(0.0534)	(0.0244)	(0.0244)	(0.0244)
Group 3	-0.0452	-0.0435	-0.0427	0.0427	0.0430	0.0430
	(0.0549)	(0.0550)	(0.0551)	(0.0229)	(0.0229)	(0.0230)

* p<0.05, ** p<0.01, *** p<0.001; $N = 771$

Table 9: Panel-Data Regressions under Trimming

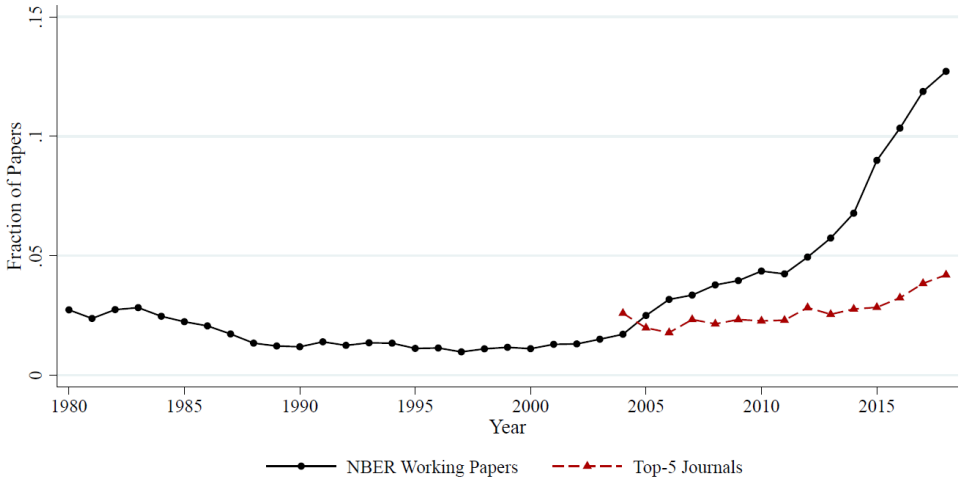
This reports regression results of the stock-day fixed-effect panel-data model with trimmed half-hour sample data for three outcomes, percentage quoted spread, consolidated displayed depth, and high-low volatility over the pre-intervention period (9 months; Jan. - Sept. 2016) and the post-intervention period (9 months; Nov. 2016 - Jul. 2017). $G1_i$, $G2_i$, and, $G3_i$ are the treatment status indicators for a pilot stock on the three treatment groups – if stock i belongs to one of the treatment groups, then the corresponding treatment group takes one and zero otherwise; $Pilot_t$ is the treatment period indicator that takes one if half-hour time index t is in the post-intervention period and zero otherwise. The columns (1) - (5) represent different choice of trimming parameters. Model (1) shows the regression results without trimming, serving as the benchmark; Model (2) to (5) contain the regression results from trimmed data, where $(\alpha, 1-\alpha)$ indicate that stocks whose ML estimate of ITE lying outside α and $1-\alpha$ percentiles are dropped out of the regression data. Fixed effects are put in place for both stocks and dates, and standard errors in parentheses are clustered by both stocks and dates as well.

	(1) No Trim	(2) (1%, 99%)	(3) (2.5%, 97.5%)	(4) (5%, 95%)	(5) (10%, 90%)
A. Percentage Qutoed Spread					
$G1_i \times Pilot_t$	17.77***	12.16***	9.82***	7.77***	5.19***
	(2.26)	(1.75)	(1.46)	(1.14)	(1.01)
$G2_i \times Pilot_t$	14.69***	10.16***	8.31***	7.09***	3.86***
	(2.37)	(1.83)	(1.34)	(1.13)	(0.99)
$G3_i \times Pilot_t$	14.68***	11.44***	9.41***	7.93***	5.42***
	(2.51)	(1.79)	(1.50)	(1.24)	(0.87)
$\bar{R}^2$	0.4471	0.4455	0.4510	0.4567	0.4636
Observations	7,622,914	7,358,294	6,965,250	6,298,848	5,069,891
B. Consolidated Displayed Depth					
$G1_i \times Pilot_t$	2154.15***	1591.95***	1146.80***	1047.53***	703.53***
	(291.28)	(156.83)	(84.79)	(65.90)	(42.96)
$G2_i \times Pilot_t$	2075.61***	1714.10***	1288.06***	997.91***	674.23***
	(247.22)	(164.31)	(106.61)	(69.01)	(40.22)
$G3_i \times Pilot_t$	3052.89***	2152.46***	1447.08***	971.01***	713.81***
	(537.77)	(210.40)	(119.14)	(68.31)	(45.01)
$\bar{R}^2$	0.4139	0.3592	0.3279	0.2921	0.2397
Observations	7,622,914	7,540,273	7,336,290	7,042,490	6,226,148
C. High-Low Volatility					
$G1_i \times Pilot_t$	-9.38***	-4.68***	-2.41*	-1.94	-0.36
	(1.61)	(1.26)	(1.13)	(1.05)	(0.91)
$G2_i \times Pilot_t$	-8.55***	-5.14***	-3.28**	-0.69	0.34
	(1.51)	(1.31)	(1.24)	(1.08)	(1.01)
$G3_i \times Pilot_t$	-6.41***	-4.33***	-2.11	-1.24	0.29
	(1.42)	(1.27)	(1.14)	(1.04)	(0.96)
$\bar{R}^2$	0.1704	0.1788	0.1815	0.1841	0.1909
Observations	7,622,914	7,314,101	6,886,315	6,147,205	4,909,678

standard errors in parentheses clustered by stocks and dates
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

B. Figures

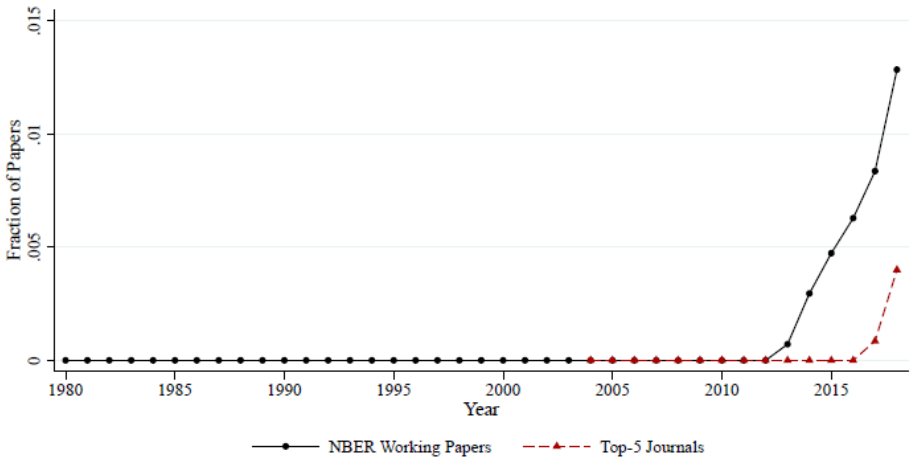
Figure 1: Randomized Controlled Trials in Applied Microeconomics



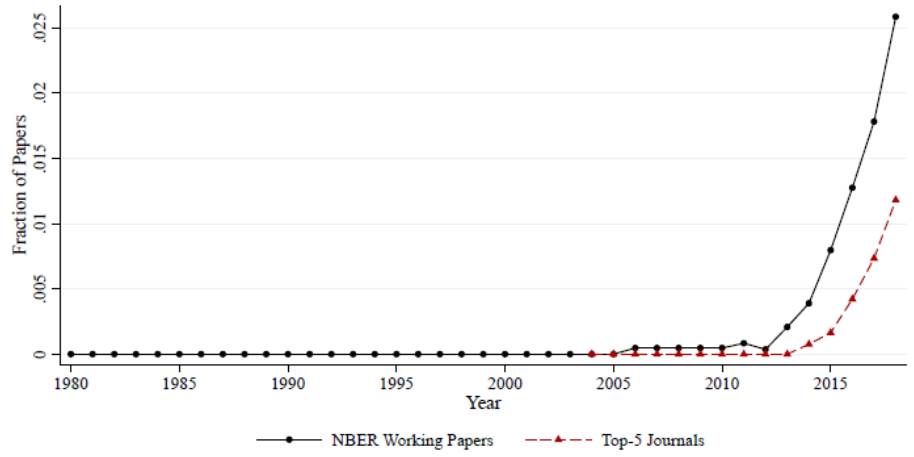
This figure, drawn from Currie, Kleven, and Zwiers (2020), shows the fraction of papers referring to randomized control trials. It counts applied microeconomics research papers among all of the National Bureau of Economic Research working papers between January 1, 1980 and June 30, 2018, and all the papers published in the top five academic journals (*American Economic Review, Econometrica, Journal of Political Economy, Quarterly Journal of Economics*, and *Review of Economic Studies*) between January 1, 2004 and August 2019.

Figure 2: Big Data and Machine Learning in Applied Microeconomics

A. Big Data



B. Machine Learning



Those figures, drawn from Currie, Kleven, and Zwiers (2020), show the fraction of papers referring to big data and machine learning. They counts applied microeconomics research papers among all of the National Bureau of Economic Research working papers between January 1, 1980 and June 30, 2018, and all the papers published in the top five academic journals (*American Economic Review, Econometrica, Journal of Political Economy, Quarterly Journal of Economics*, and *Review of Economic Studies*) between January 1, 2004 and August 2019.

Figure 3: Sample Periods

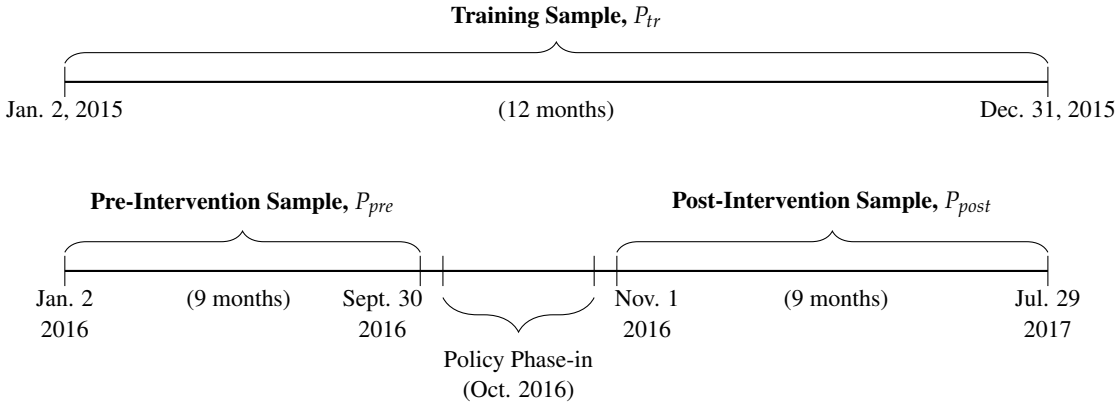
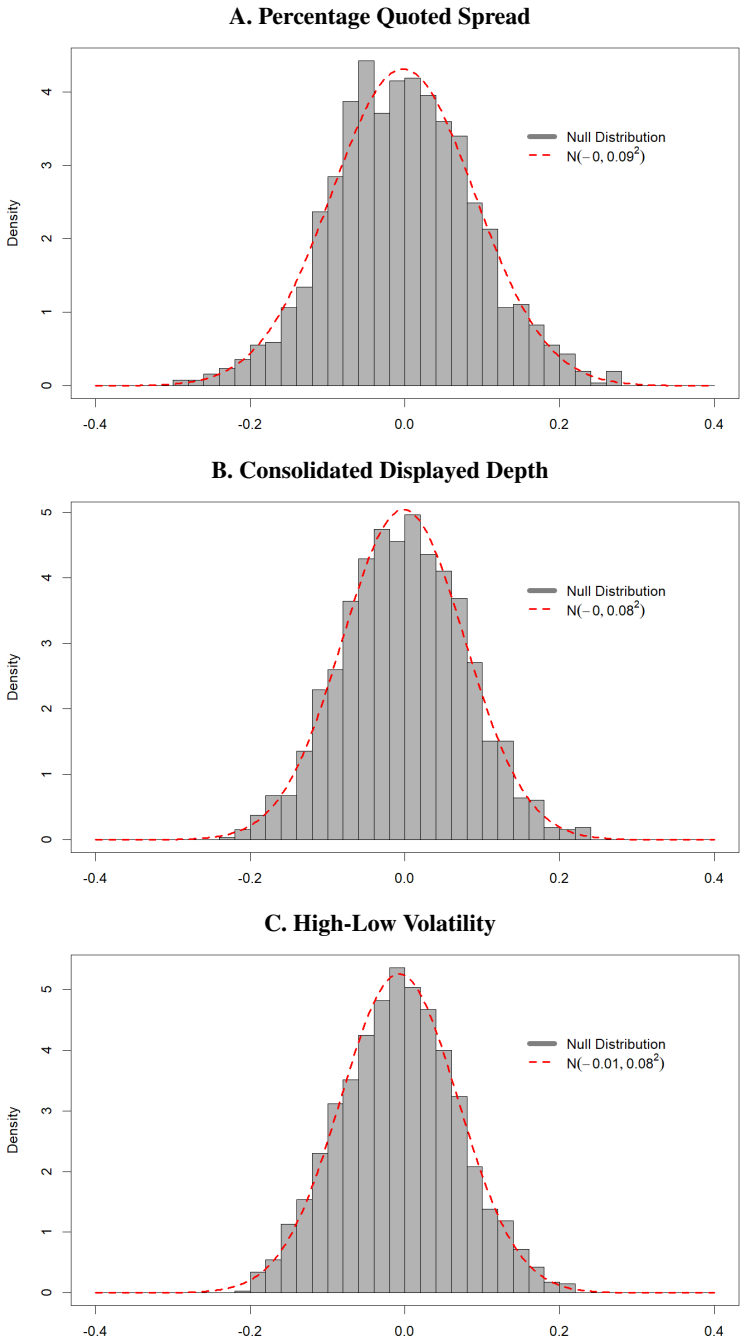
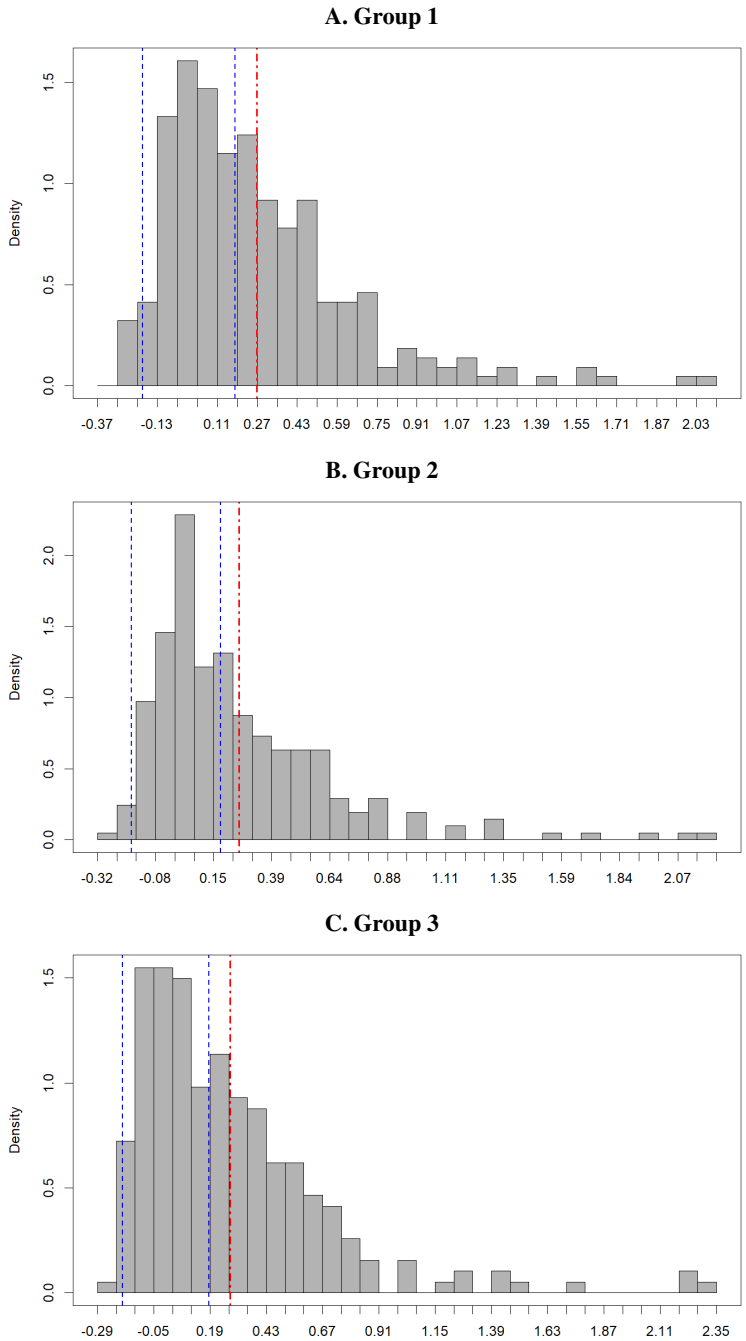


Figure 4: The Null Distributions



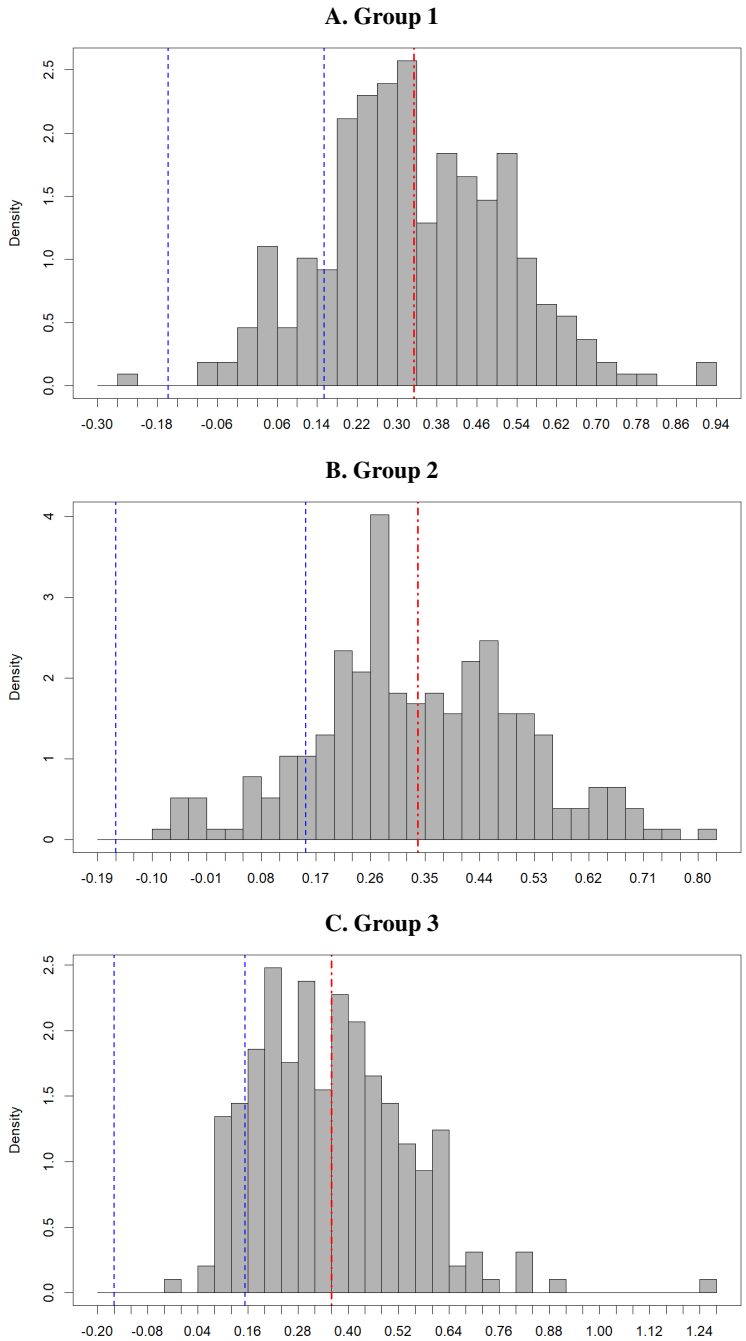
The figures show the distributions of the ITE test statistics, defined in (9), for the stocks in the donor pool, where the lower and upper 0.5% of the individual estimates are trimmed out beforehand. The histograms drawn by the gray bars represent the null distribution for each of percentage quoted spread, consolidated displayed depth, and high-low volatility. The red dotted lines are the normal distributions whose mean and variance are obtained from the individual estimates.

Figure 5: Distribution of ITEs: Percentage Quoted Spread



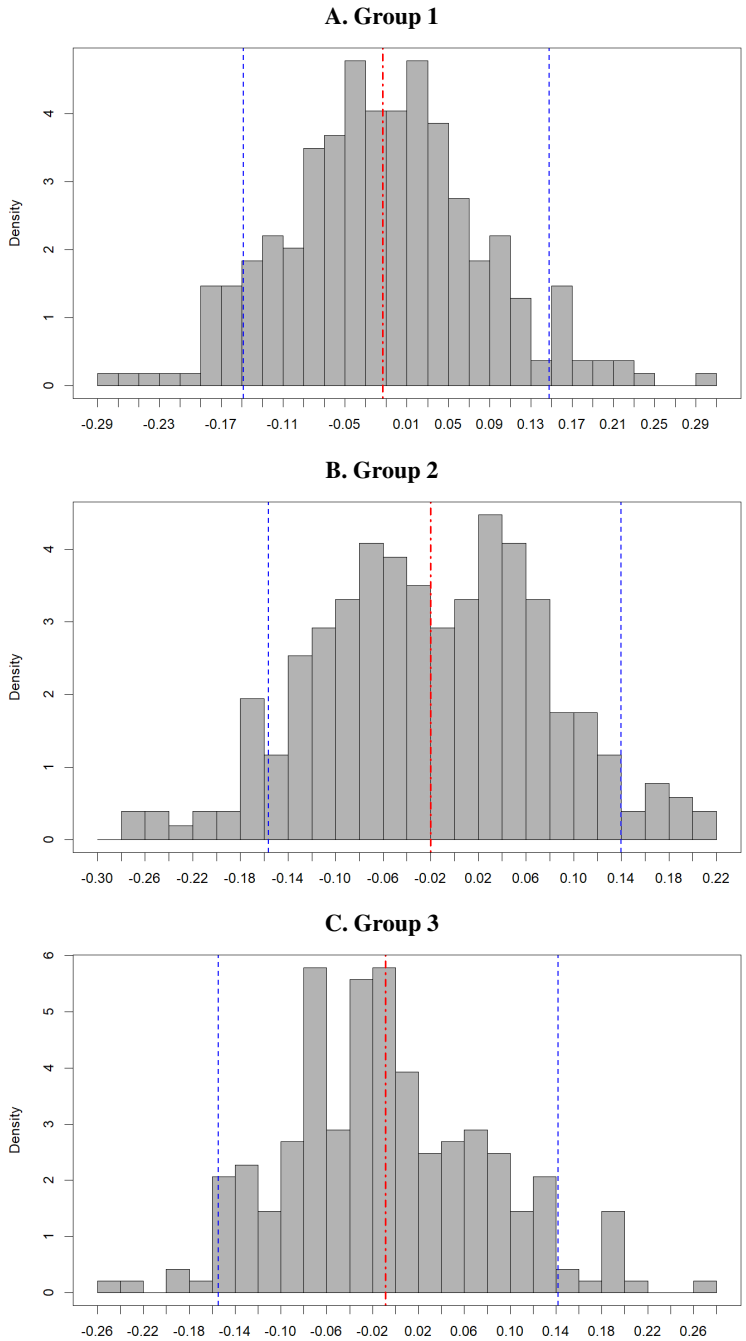
The figures show the distributions of the ITE test statistics, defined in (9), for percentage quoted spread for the sample stocks in the three treatment groups. The histograms drawn by the gray bars represent the distributions of the ML estimates. The blue dotted vertical lines indicate critical values at the two-sided 5% significance level, computed from the corresponding null distributions in Figure 4.

Figure 6: Distribution of ITEs: Consolidated Displayed Depth



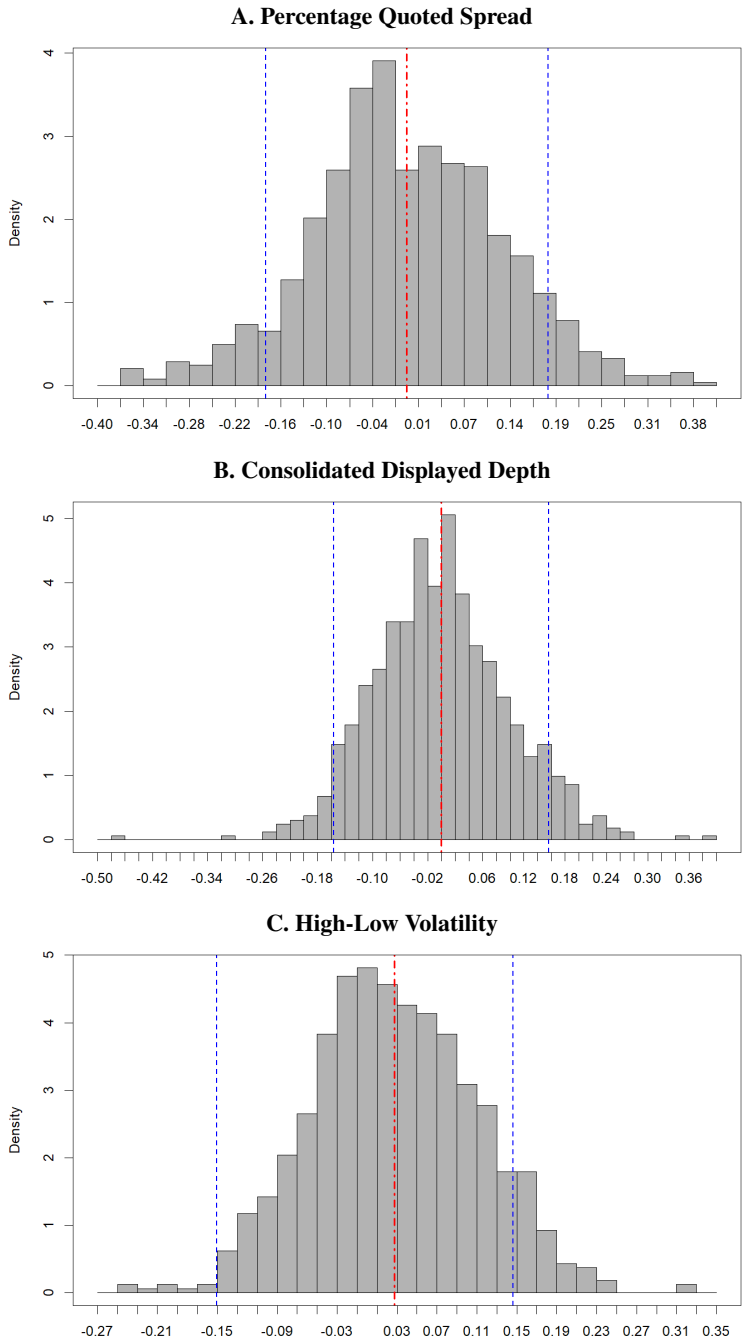
The figures show the distributions of the ITE test statistics, defined in (9), for consolidated displayed depth for the sample stocks in the three treatment groups. The histograms drawn by the gray bars represent the distributions of the ML estimates. The blue dotted vertical lines indicate critical values at the two-sided 5% significance level, computed from the corresponding null distributions in Figure 4.

Figure 7: Distribution of ITEs: High-Low Volatility



The figures show the distributions of the ITE test statistics, defined in (9), for high-low volatility for the sample stocks in the three treatment groups. The histograms drawn by the gray bars represent the distributions of the ML estimates. The blue dotted vertical lines indicate critical values at the two-sided 5% significance level, computed from the corresponding null distributions in Figure 4.

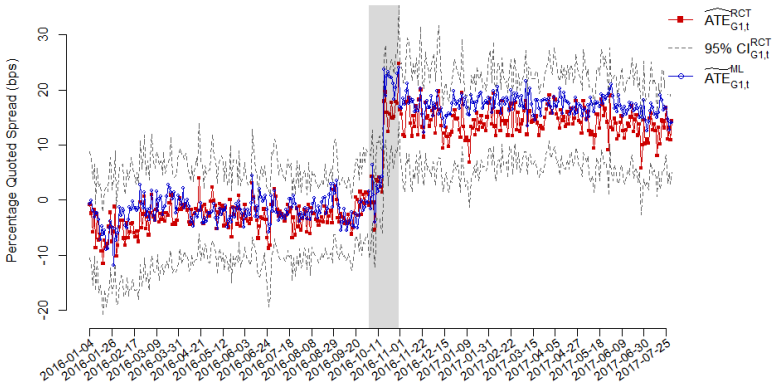
Figure 8: Distribution of Individual Spillover Effects



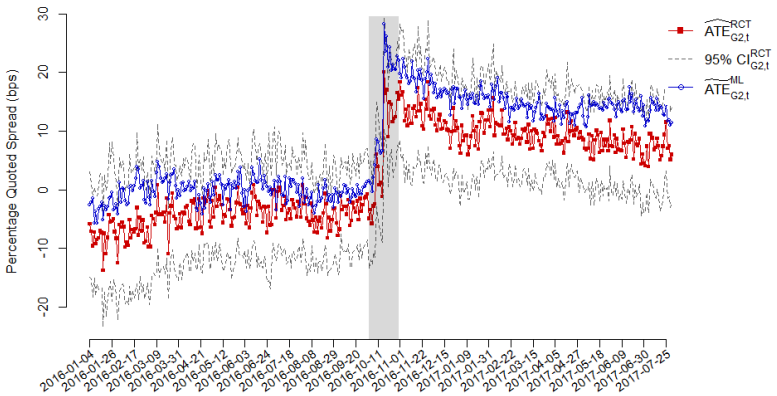
The figures show the distributions of the ITE test statistics, defined in (9), for the three outcomes, percentage quoted spread, consolidated displayed depth, and high-low volatility, for the sample stocks in the control group. For each outcome, the histograms drawn by the gray bars represent the distributions of the ML estimates.. The blue dotted vertical lines indicate critical values at the two-sided 5% significance level, computed from the corresponding null distributions in Figure 4.

Figure 9: Time-Series of ATE Estimates: Percentage Quoted Spread

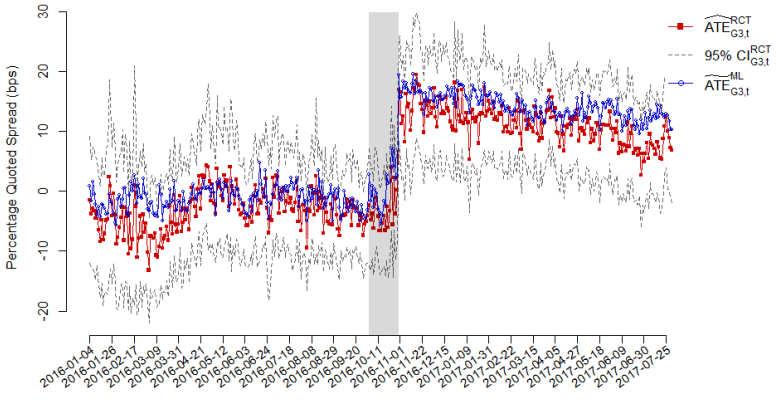
A. Group 1



B. Group 2



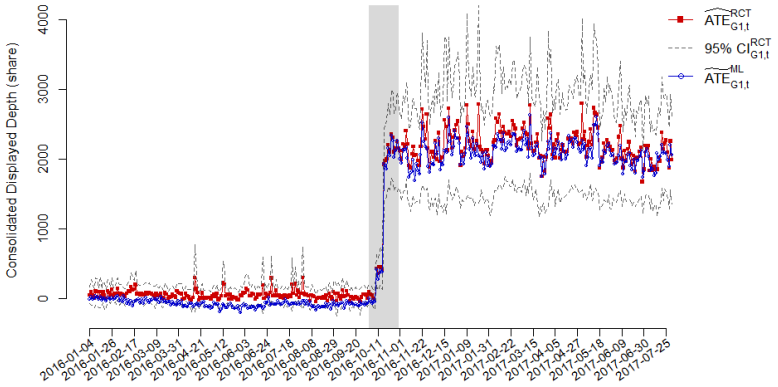
B. Group 3



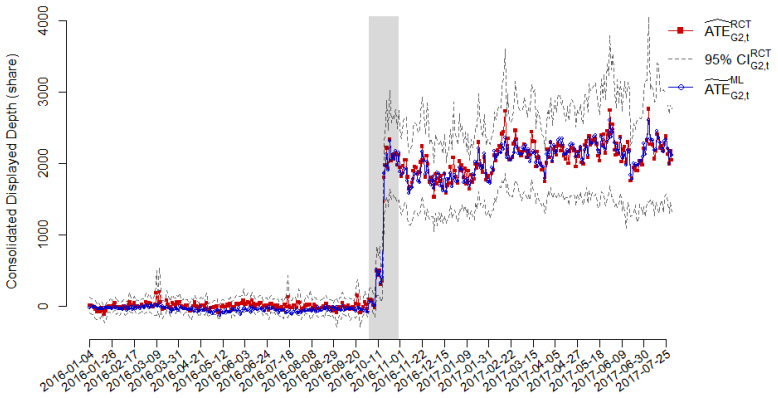
The figures show daily time series of cross-sectional average treatment effects for each treatment group in TSPP. The red solid lines show the RCT estimates (14) and gray dotted lines indicate their 95% confidence intervals on variance estimates (15) under normal approximation. The blue solid lines represent the ML estimates (16), which do not employ a control group in the conventional sense.

Figure 10: Time-Series of ATE Estimates: Consolidated Displayed Depth

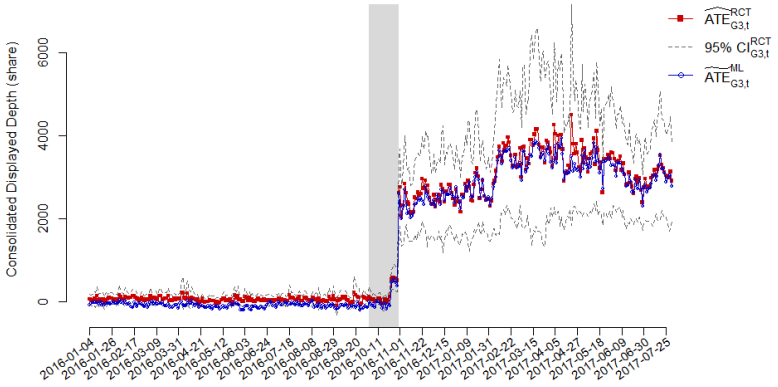
A. Group 1



B. Group 2



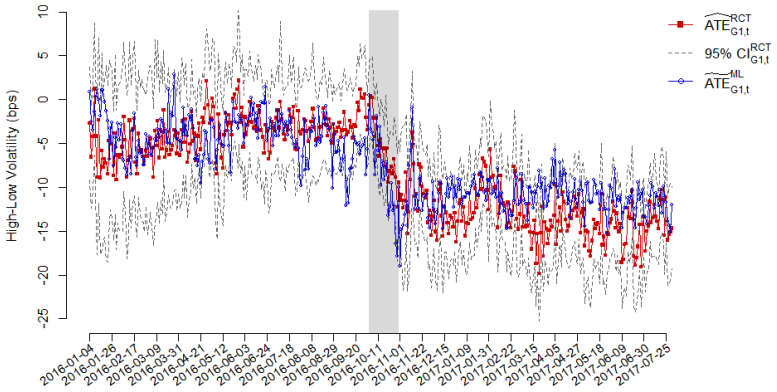
B. Group 3



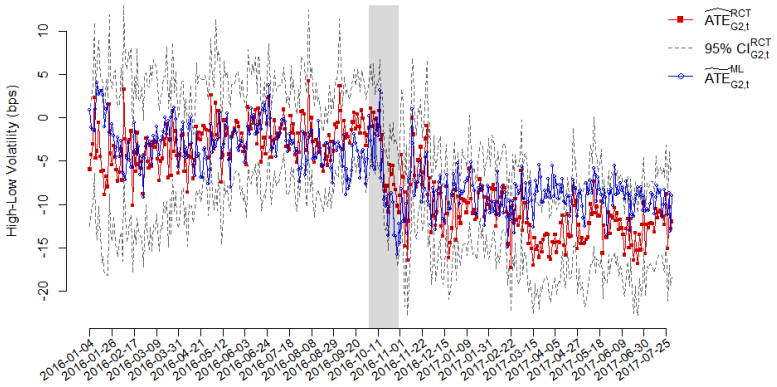
The figures show daily time series of cross-sectional average treatment effects for each treatment group in TSPP. The red solid lines show the RCT estimates (14) and gray dotted lines indicate their 95% confidence intervals on variance estimates (15) under normal approximation. The blue solid lines represent the ML estimates (16), which do not employ a control group in the conventional sense.

Figure 11: Time-Series of ATE Estimates: High-Low Volatility

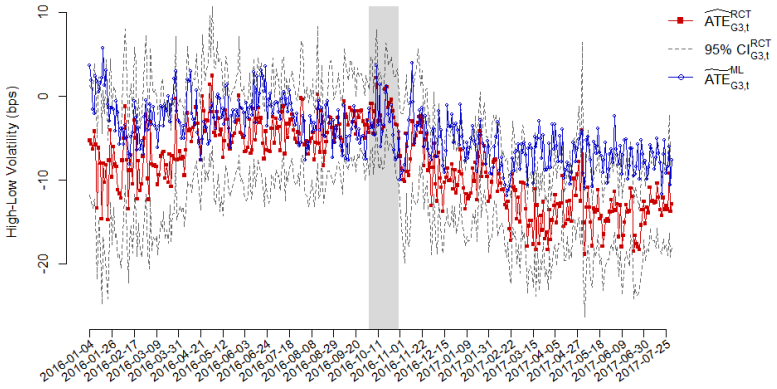
A. Group 1



B. Group 2



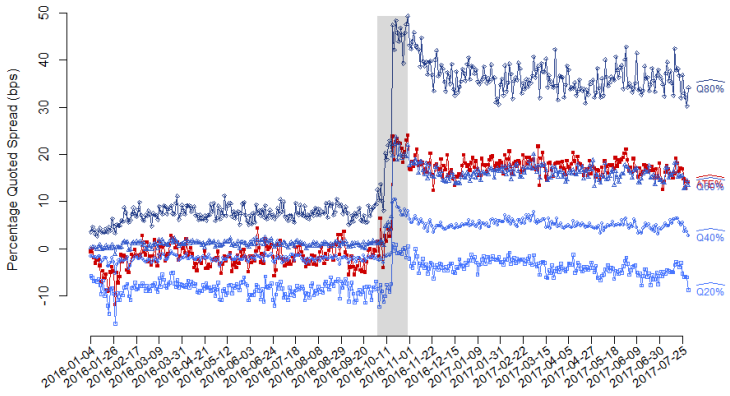
B. Group 3



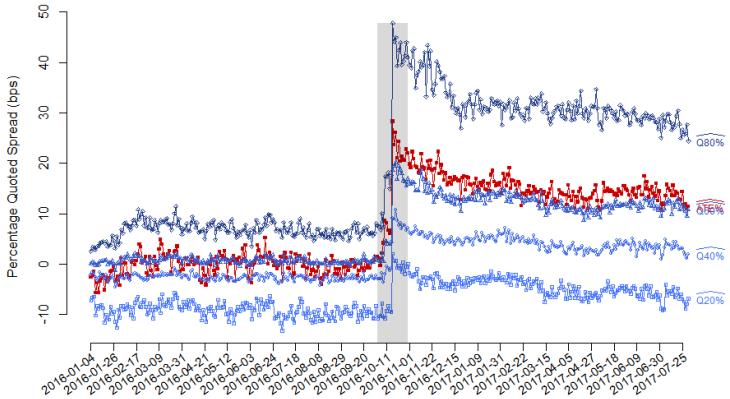
The figures show daily time series of cross-sectional average treatment effects for each treatment group in TSPP. The red solid lines show the RCT estimates (14) and gray dotted lines indicate their 95% confidence intervals on variance estimates (15) under normal approximation. The blue solid lines represent the ML estimates (16), which do not employ a control group in the conventional sense.

Figure 12: Time-Series of Quantiles: Percentage Quoted Spread

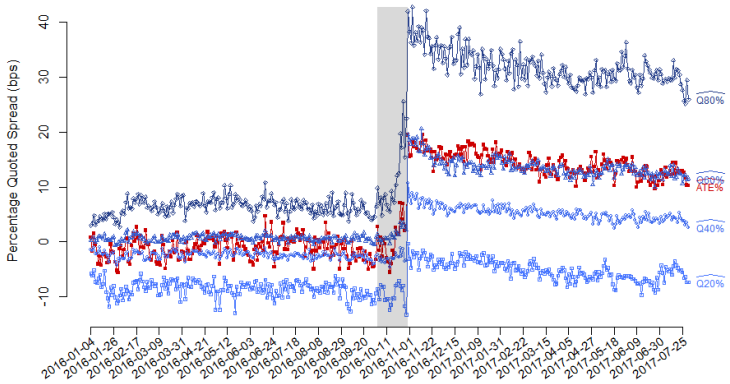
A. Group 1



B. Group 2



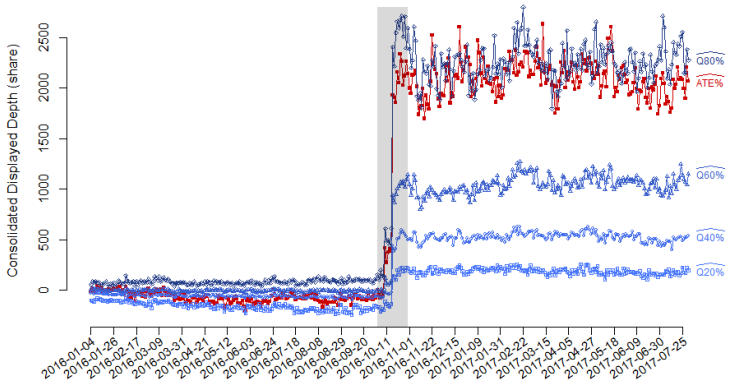
B. Group 3



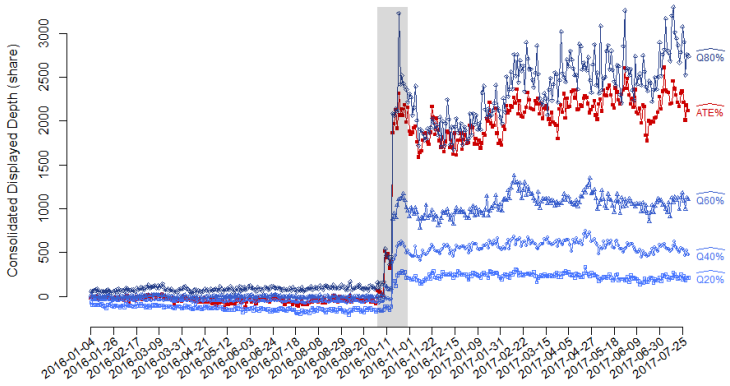
The figures show daily time-series of cross-sectional quantile policy effects by the treatment groups. The policy-effect estimates on each group based on the ML prediction errors, $\{\hat{\Delta}_{i,t}\}_{i\in G}$ in (16), are computed day-by-day. The bluish lines represent quantile values at 80%, 60%, 40% and 20% from top to bottom. The red line draws daily average values.

Figure 13: Time-Series of Quantiles: Consolidated Displayed Depth

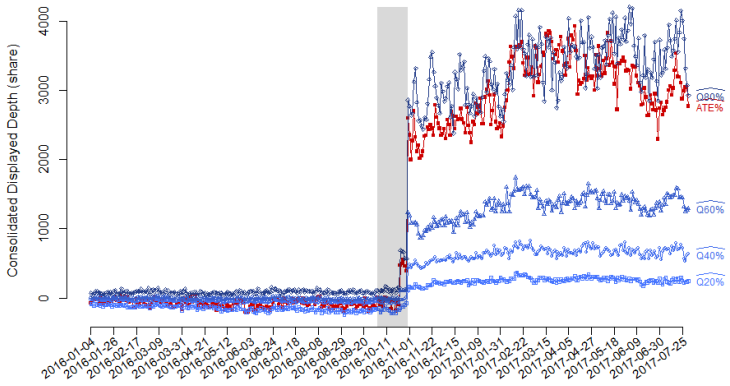
A. Group 1



B. Group 2



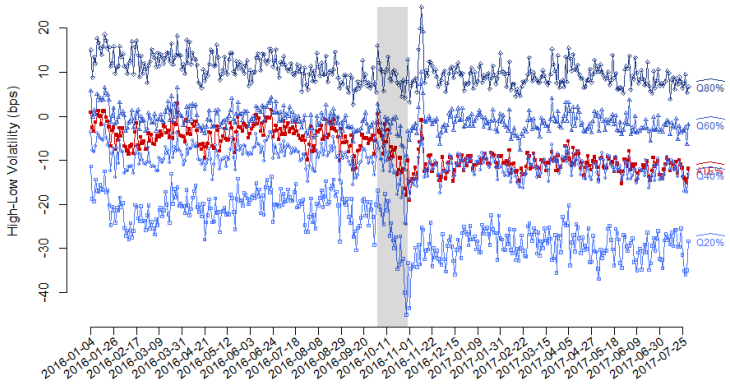
B. Group 3



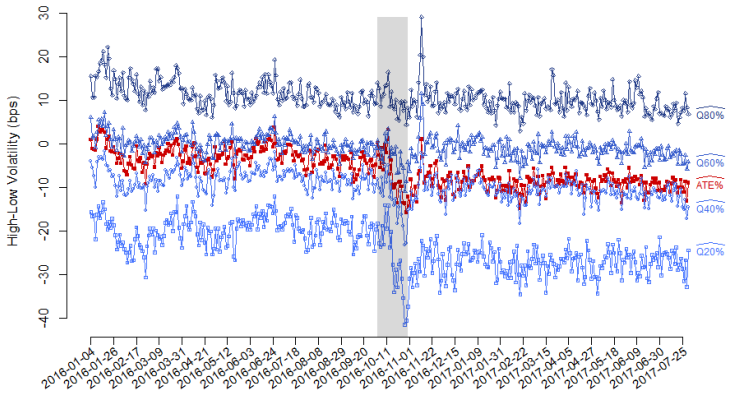
The figures show daily time-series of cross-sectional quantile policy effects by the treatment groups. The policy-effect estimates on each group based on the ML prediction errors, $\{\hat{\Delta}_{i,t}\}_{i\in G}$ in (16), are computed day-by-day. The bluish lines represent quantile values at 80%, 60%, 40% and 20% from top to bottom. The red line draws daily average values

Figure 14: Time-Series of Quantiles: High-Low Volatility

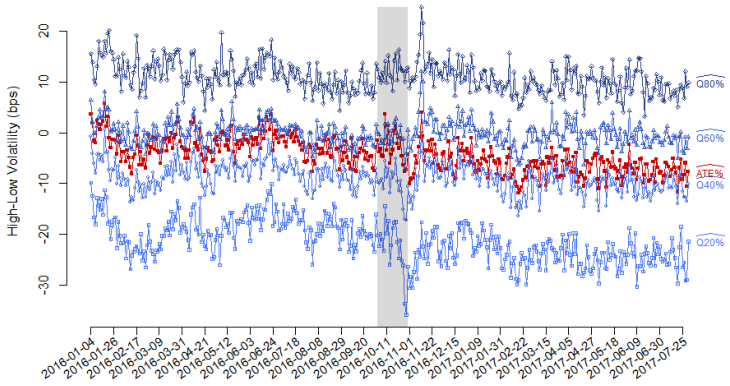
A. Group 1



B. Group 2



B. Group 3



The figures show daily time-series of cross-sectional quantile policy effects by the treatment groups. The policy-effect estimates on each group based on the ML prediction errors, $\{\hat{\Delta}_{i,t}\}_{i\in G}$ in (16), are computed day-by-day. The bluish lines represent quantile values at 80%, 60%, 40% and 20% from top to bottom. The red line draws daily average values

C. Technical Details in Data Processing

C.1. The Half-Hour Version of National Best Bid and Offer (NBBO)

Daily TAQ quote data simply show updates of best quotes on individual exchanges in chronological order. That said, they do not immediately deliver NBBO, dynamically adjusted upon arrival of the new updates across local markets in the U.S. Thus, the use of Daily TAQ quote data requires running a sorting process to produce NBBO at each quote update in Daily TAQ quote data. The core principle in the NBBO processing is to find the maximum (minimum) among the locally highest (lowest) bids (offers) at each second of the quote updates in the data. I conduct this NBBO processing for all the U.S.-traded stocks over the whole sample period with the filtering rules drawn from the HJ algorithm. After that, I perform time-averaging on the NBBO-sorted quote data for each stock on each trading day over every half-hour interval during the regular training session, excluding the first and last five minutes, i.e., the first and last half-hour intervals span over 9:35 - 10:00 AM and 3:30 - 3:55 PM rather than 9:30 - 10:00 AM and 3:30 - 4:00 PM, respectively.[59] In doing so, I also compute the highest and lowest midquotes of NBBO on each half-hour interval.

Importantly, there have been a few occasions at the stocks-days level that the first *valid* half-hour interval is not the one starting from 9:35 AM. In principle, all the stocks are expected to have the full 13 half-hour intervals on any given trading day between 9:35 AM and 3:55 PM. Because of the nature of the quote data, if there is an initial update around the opening bell, the corresponding stock must have the full 13 intervals, regardless of whether there is a trade executed or not. Though, this is not always case because the HJ algorithm filters out certain quotes, possibly placed around the opening bell, considered as wrongly recorded, deleting observations in raw Daily TAQ quote data and so making some stocks on certain days have the first valid half-hour interval from the second interval, i.e., the half-hour interval starting from 10:00 AM. However, these are rare incidents, and I only include stocks-days that have the first valid half-hour interval at latest starting from the second to exclude late starters.

There are a couple of remarks on this data processing. First, NBBO is a dynamic concept. Unlike to trades which are a static concept only meaningful at the time of the events, a NBBO update is considered as valid until a new update arrives. Thus, time-averaging over a predetermined length of time does not erase the original interpretation of NBBO while offering readily a synchronized data structure in panel data. Secondly, compared to the prior works that perform time-averaging over full trading days and so average out all the potentially interesting intraday variations, this paper keeps intraday activities to a meaningful degree, taking richer information into consideration in dealing with the similar research questions. Importantly, the use of the intraday frequency is reflective of the literature that has long looked into intraday patterns on liquidity and volatility measures (e.g., Chung, Van Ness, and Van Ness (1999); Foster and Viswanathan (1993); Lin, Sanger, and Booth (1995); Madhavan, Richardson, and Roomans (1997); McInish and Wood (1992); Stoll and Whaley (1990)). In addition, the information set of the half-hour time-averaged data nests that of the daily time-averaged data given that an averaged outcome over the half-hour intervals on a trading day are the daily version on that day.

[59] Dropping the first and last five minutes is to avoid potential contamination from the call auction process around the opening and closing bells

C.2. Stationarity Test for Construction of Donor pools

For a given stock, I first trim out the time series of its outcome at the lower and upper 1% quantiles over the pre- and post-intervention sessions separately. This is to avoid the impact of outliers. Also, before running the tests, I conduct a preliminary filtering on the time-series by taking orthogonal projection. This is to reflect the possibility that certain portions of the time series variation are driven by some known factors that are apparently exogenous of TSPP. If the stationarity tests are performed without factoring out those variations, they can mistakenly rule out the stock that fails to survive in the tests because of the variations induced by the known factors but not presence of intervened effects from TSPP. For it, I regress the time series on VIX and half-hour interval fixed effects dummy variables that approximate such known factors, and take the residuals. Finally, I run stationarity tests on the residuals for each stock, where stationarity tests include a random walk test in the pre-intervention period, a long-run mean difference test between the pre- and post-intervention periods, and structural break tests on an autoregressive model between the pre- and post-intervention periods.

The random-walk test is based on the simplest specification including only the lagged term. If the time-series follows a random walk in the pre-intervention periods, it cannot serve as a valid predictor in the proposed ML procedure. The two-sample long-run mean t-test looks at the standard t-value on the difference of the time-series sample means between the pre- and post-intervention periods divided by the standard errors, where the standard errors are computed from the Newey-West variance estimator with a well-known rule-of-thumb truncation parameter $m_k = 0.75 \times T_k^{1/3}$, $k \in \{pre, post\}$ in place. If the long-run means between the two periods are highly different, then it may signal that the outcome of the stock is significantly affected by the policy intervention under TSPP. Lastly, the structural break test examines possible breaks on the dynamics of the outcome across the pre- and post-intervention periods. For it, I fit the residuals obtained from the prelininary regression to the AR (38) model for the pre- and post- intervention samples, where the 38-lag choice reflects potential persistence of the information up the three consecutive trading days, 13 intervals $\times$ 3 days $-$ 1 lag $=$ 38. Then, I construct a Wald-statistics excluding the constant term, and carry out the Chi-square test for the potential structural break between the two periods.[60]

With all the three test results for each outcome at hand, I first exclude the stocks on which I cannot reject at the 5% signficance level the null that the time-series of their outcomes follows a random walk. Among those that reject the null, I further exclude the stocks that show statistical significance at the 1% level on the both long-run mean differences and structural break tests. Subsequently, the donor pool for each of the three outcomes, percentage quoted spread, consolidated displayed depth, and high-low volatility, in order has 1,280, 1,343 and 2,045 member stocks.

[60] Bruce E. Hansen provides an excellent summary of popular structural break tests in the time-series literature, including the Chi-squared-based test I adopted in this paper. The related slides are available on his website at https://www.ssc.wisc.edu/∼bhansen/crete/crete5.pdf.