Online Appendix for
The Value of Last-Mile Delivery in Online Retail

Zhikun Lu
Shanghai New York University
Ruomeng Cui
Goizueta Business School, Emory University
Tianshu Sun
Cheung Kong Graduate School of Business
Lixia Wu
Cainiao Network

A. Additional Figures and Tables

A.1. Hypotheses

Our hypotheses are summarized in Figure A1.

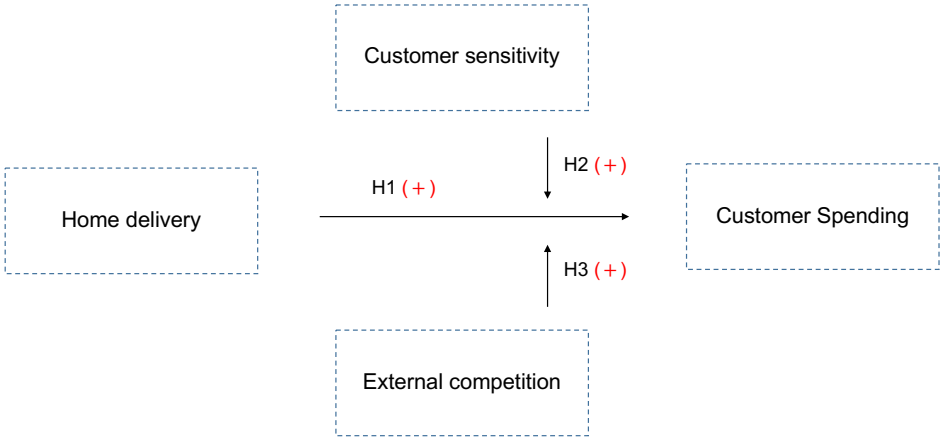


Figure A1 Hypothesis Framework

A.2. A Cainiao Station

Figure A2 shows the typical location of a Cainiao Station and the view of the store front.

A.3. Voice from Customers

Cainiao's original strategy was focusing exclusively on self-pickup. However, since customer preferences are intrinsically diverse, this approach did not cater to everyone. Two contrasting customer

(a) Location (b) Front view

Figure A2 Cainiao Station as Local Pickup Center

reviews about Cainiao Station highlight this diversity. As shown in Figure A3, many customers value the pickup mode, as evident from the positive review and high rating. Indeed, Cainiao Stations provide a secure and flexible method for package reception. On the other hand, the absence of doorstep home delivery is seen as a significant drawback, as indicated by the negative review and low rating in Figure A4. For many, home delivery is an indispensable convenience. While Cainiao's initial emphasis on self-pickup was efficient, it inadvertently neglected a portion of their customers who favored home delivery.

The contrasting customer reviews convey two messages. First, the absence of home delivery is a significant concern for some customers. In fact, this was exactly the motivation for Cainiao to launch the home delivery program. Second, the distinct reactions to Cainiao's pickup service highlight the heterogeneity in customer preferences. While some find the station-based pickup convenient, others clearly miss the comfort of receiving packages at their doorstep. This disparity suggests the opportunity for Cainiao to further tailor its home delivery services, catering to the diverse needs of its customer base without escalating costs.

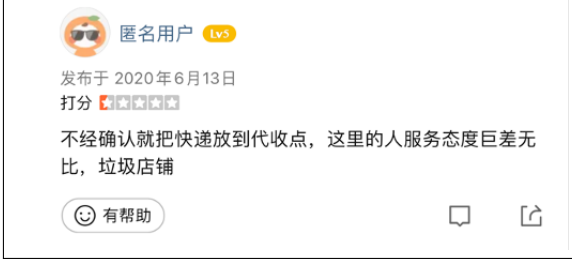
A.4. Geographic Coverage and City Classification

Our sample encompasses 284 out of 333 prefecture-level divisions (cities) in mainland China, achieving 85.3% geographic coverage. The actual share of orders covered would be even much higher, as the excluded cities are typically smaller and less economically active. We classify cities using Alibaba's internal city-tier system, which aligns closely with established media rankings and captures contemporary patterns of economic development and market sophistication. Table A1 presents the classification scheme and the distribution of observations across city tiers.

Anonymous User
Date: June 24, 2020
Rating: 5 (highest)
Review: Cainiao Station right next to my home. Very convenient for either shipping or picking up parcels. Great customer service. Everything is well organized on the pickup rack.

Comment by another user: They used to deliver packages to the doorstep. With this (i.e., Cainiao Station), now I have to pick up myself every time. How can it be convenient! I'm one kilometer away and the trip is so annoying.

Figure A3 An Example of Positive Customer Review



Anonymous User
Date: June 13, 2020
Rating: 0.5 (lowest)
Review: Left my packages at the Cainiao Station without my consent. Terrible customer service.

Figure A4 An Example of Positive Customer Review

B. Robustness Analysis

This appendix presents a comprehensive set of robustness analyses to validate the credibility of our findings. We begin by estimating dynamic treatment effects and testing the parallel trends assumption (Section B.1), followed by a placebo test for additional validation (Section B.2). To further address potential endogeneity concerns, we combine DiD with matching in Section B.3. We then assess whether our results are driven by time-varying local economic or social factors by controlling for city- and station- specific trends (Section B.4), and we explore the potential confounding effects of the COVID-19 pandemic (Section B.5). We also test for potential spillover effects stemming from untreated customers' perceptions through an additional placebo analysis and find no significant evidence (Section B.6). In Section B.7, we adopt the classic DiD framework and simulate a scenario where all customers from the treatment group receive treatment simultaneously. Recognizing the disproportionate influence of high-frequency customers, we re-estimate our models using weighted regressions based on customer order volume (Section B.8). Moreover, we test the robustness of our results to alternative outcome transformations, including $\log(x+0.01)$ and $\text{arcsinh}(x)$ (Section B.9). Finally, we replicate the main analysis using both raw (unlogged)

Table A1 City Tier Classification

Tier	Cities	N
Tier 1	Beijing (北京市), Shanghai (上海市), Guangzhou (广州市), Shenzhen (深圳市), Hangzhou (杭州市)	8,850
Tier 2	Baotou (包头市), Changchun (长春市), Changsha (长沙市), Changzhou (常州市), Chengdu (成都市), Chongqing (重庆市), Dalian (大连市), Daqing (大庆市), Dongguan (东莞市), Foshan (佛山市), Fuzhou (福州市), Guilin (桂林市), Guiyang (贵阳市), Harbin (哈尔滨市), Hefei (合肥市), Hohhot (呼和浩特市), Huai'an (淮安市), Huizhou (惠州市), Jiaxing (嘉兴市), Kunming (昆明市), Linyi (临沂市), Luoyang (洛阳市), Nanchang (南昌市), Nanjing (南京市), Nanning (南宁市), Nantong (南通市), Ningbo (宁波市), Qingdao (青岛市), Quanzhou (泉州市), Shenyang (沈阳市), Shijiazhuang (石家庄市), Suzhou (苏州市), Taiyuan (太原市), Tianjin (天津市), Weifang (潍坊市), Weihai (威海市), Wenzhou (温州市), Wuhan (武汉市), Wuxi (无锡市), Xiamen (厦门市), Xi'an (西安市), Xianyang (咸阳市), Xuzhou (徐州市), Yangzhou (扬州市), Yantai (烟台市), Zhengzhou (郑州市), Zhenjiang (镇江市), Zhongshan (中山市), Zhuhai (珠海市)	65,297
Tier 3	All other cities not listed above	25,853

Notes: City classification follows conventional economic development and administrative importance criteria. Tier 1 represents the most economically developed mega-cities with national significance. Tier 2 includes major provincial capitals and economically important cities. Tier 3 includes all remaining cities in the sample. N denotes the number of observations from each city tier.

and winsorized measures of spending and average order value (Section B.10). All these robustness checks attest to the stability and validity of our findings.

B.1. Dynamic Treatment Effects and Parallel Trends

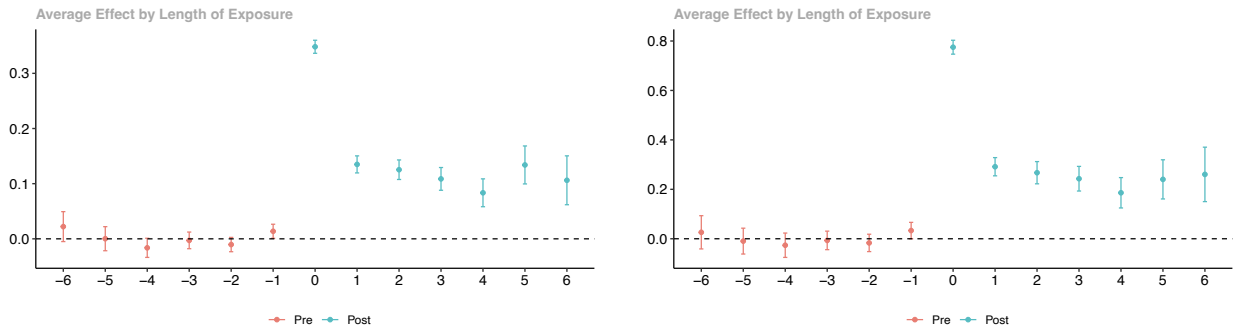
In this section, we estimate the dynamic treatment effects of home delivery, which not only shows the treatment effects' magnitude over time but also tests the parallel trends assumption under the staggered adoption setting. To do this, researchers often use TWFE regressions that include leads and lags of the treatment. An exemplary specification is illustrated by as follows:

$$y_{it} = \alpha_i + \sum_{s=-K}^{-2} \tau_s D_{it}^s + \sum_{s=0}^{L} \tau_s D_{it}^s + \theta_t + e_{it}, \tag{7}$$

which is a "dynamic" version of Equation (1). Here, K is the maximum lag period, L is the maximum lead period, D_{it}^l is an indicator which equals 1 only if unit i is from the treatment group and received treatment in period $t-l$. Therefore, when $s \geq 0$, τ_s captures the dynamic post treatment effects. On the other hand, when $s < 0$, τ_s captures the pre-treatment effects and is expected to be zero under the no-anticipation and the parallel trends assumptions, which forms the basis of parallel trends test (Sun and Abraham 2021).

Recent studies (e.g., Callaway and Sant'Anna 2021, Sun and Abraham 2021) highlight several issues with TWFE regressions in staggered DiD setting when treatment effect heterogeneity exists,

a scenario plausible in our case. Specifically, parameters like τ_s can be decomposed as a weighted averages of individual-level treatment effects. Yet, TWFE estimators often assign counterintuitive weights, including negative ones, which can bias the estimates and even reverse the sign of τ_s. Sun and Abraham (2021) further demonstrates that pre-trend tests based on pre-treatment coefficients can be invalid due to the contamination issue caused by treatment effect heterogeneity. The root cause of negative weighting and contamination is that TWFE makes "forbidden comparisons" between early-treated and later-treated units. To address this problem, we estimate dynamic treatment effects using the method proposed by Callaway and Sant'Anna (2021), which avoids such comparisons and adopts a more refined weighting scheme.



(a) Log Orders (b) Log Spending

Figure B1 Dynamic Treatment Effects under Staggered Adoption

The estimation results are visualized in Figure B1, which shows the size of treatment effects six months before and after the treatment. We find that:

1. customers in the treatment group did not increase their spending before they received home delivery, which verifies the parallel trends assumption;

2. the magnitude of treatment effects becomes stable in the long term;

3. the magnitude of treatment effects is consistent with the estimation results from the TWFE regression models in Table 2.

B.2. Placebo Test

Placebo tests are widely used in DiD analysis to examine the robustness of estimation results. The main idea is to check whether any treatment effect can be detected on a fake treatment group. Depending on how the fake treatment group is constructed, placebo tests may take different forms. A popular placebo test under the classic DiD (i.e., with homogeneous treatment timing) is to assume the treatment is implemented before the actual implementation date. Since such treatment is fake, one should not be able to observe any effect before the treatment is actually implemented.

We design our placebo test following the above idea. Although the customers in our study received home delivery in different periods, we can still find a subgroup who received the treatment in the same period. Specifically, we select the cohort that received home delivery in August, 2021 (as shown in Figure 2, this cohort is large so that we have more treated units for test and is early in timing so that we have more treated periods for observation) as our treatment group but assume their received home delivery in April, 2021. Consequently, we expect no impact being observed between April and July, 2021 and true impact emerging exactly in August.

We perform the test using the synthetic difference-in-differences (SDiD) method proposed by Arkhangelsky et al. (2021). The main advantage of this method is that it automatically matches the pre-treatment trends between the treatment and control groups by constructing a synthetic control group through reweighting. Figure B2 visualizes our estimation results, showing the time trends of log orders for both groups. Since the fake treatment is imposed in April (which corresponds to T=16 in Figure B2), SDiD constructs a control group whose pre-April purchasing behavior closely resembles that of the treatment group. We can see that SDiD does an excellent job at matching the pre-treatment trends as the two lines almost perfect aligned before April. Crucially, we find no significant treatment effects after the fake treatment. The average treatment effect from April to July is 0.012 with a 95% confidence interval of [-0.003, 0.027]. However, when the real treatment hit in August (which corresponds to T=20 in Figure B2), the orders from the treatment group immediately jumped up compared to the control group. Further, we find that the impact of home delivery is persistent and stable over time, and their magnitude aligns with our previous estimates in Section 5 and Section B.1. The consistency of the impact's timing and magnitude suggests that our findings are robust.

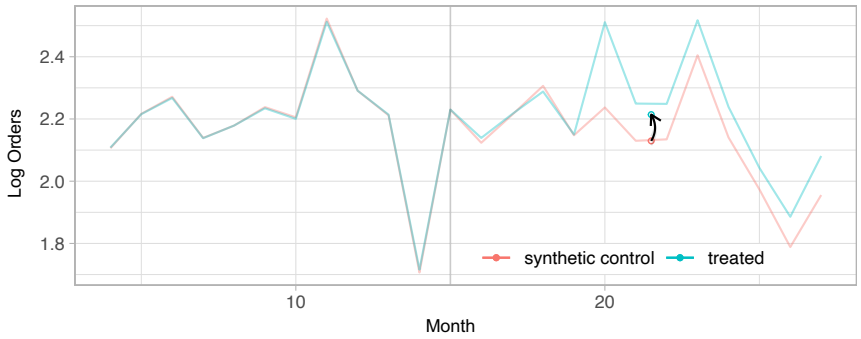


Figure B2 Placebo Test on Log Orders

B.3. DiD with Matching

Since Cainiao Stations' decisions to join the home delivery program is not random, their customers can also be systematically different. This is not an issue to DiD because the individual fixed effects eliminate all time-invariant confounders and the parallel trends assumption rules out all time-invariant confounders. Therefore, nonrandom assignment may influence our interpretation of the causal estimates (i.e., ATT) but does not compromise our identification. Nonetheless, the literature often employs matching to achieve a more balanced sample for subsequent DiD analyses (e.g., Fisher et al. 2019, Cui et al. 2020). Such an approach enables a more direct "apples-to-apples" comparison, thereby enhancing the credibility of empirical findings.

We perform one-to-one matching with replacement based on the Mahalanobis distance. We find customers' spending patterns are mostly related to their order volume and its growth rate. Therefore, we calculate each customer's pre-treatment order volume (specifically, the total number of orders in fiscal year 2021) and the corresponding annual growth rate, using these two metrics for matching.[20] To reduce regional imbalance, the matching is performed by city: a customer from the treatment group is paired with a "twin" customer from the control group only if they are located in the same city.[21] In this way, a metropolitan customer in Shanghai will be paired with another customer from Shanghai, not someone from rural areas, which helps address concerns raised in Section B.4.

We plot the time series of order volume using the matched sample in Figure B4. The number of orders typically follows a fat-tailed distribution, making it noisier than outcomes after log transformation. Moreover, we do not directly match the time paths of the two groups; rather, we only match the overall volume and annual growth rate. Despite the noise nature and indirect approach, the two monthly time series align remarkably well. We conducted a t-test for each month and found no statistical difference between the groups before the home delivery program (See Figure B3).

Finally, we estimate Equation (1) using the matched samples. As we can see, the estimation results in Table B1 are highly consistent with those without matching in Table 2, suggesting our finds are robust.

B.4. Control for Local Trends

Control for City Trends. One might be concerned that the local social and economic conditions could drive both customer spending and the adoption or expansion of home delivery service. Consequently, our baseline results in Table 2 may suffer from omitted variable bias. To address this

[20] Adding more customers features does not change much the subsequent regression results but reduces the matching quality, as finding a similar object in a higher dimensional space becomes much more challenging.

[21] Our matching caliper is set at 0.05, a threshold lower than commonly employed values like 0.1 or 0.2. A tighter caliper improves matching quality but may substantially reduces the sample size as well as the representativeness of a subsample. Given our large sample size, we prioritize higher quality matches with a smaller caliper; representativeness is less a concern as long as the estimates remain consistent.

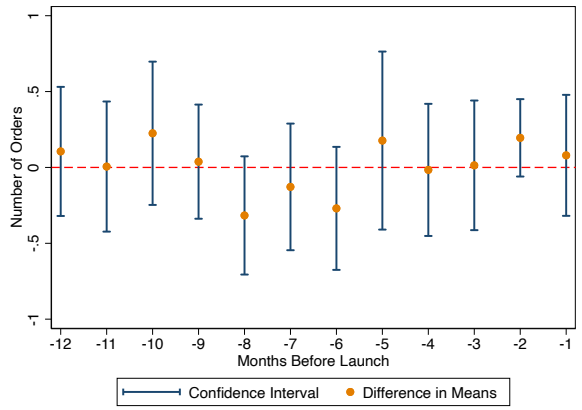


Figure B3 **Pre-Treatment Group Differences After Matching**

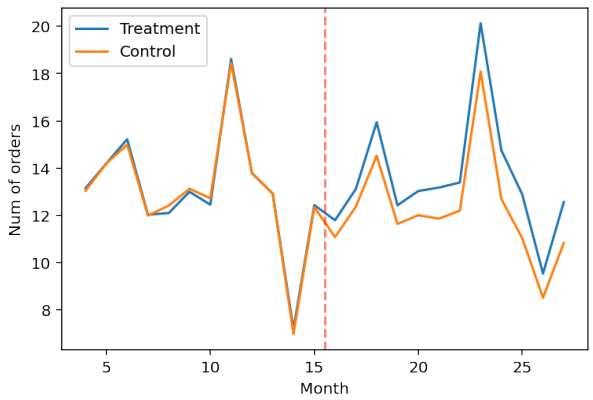


Figure B4 **Trends after Matching**

Table B1 **Impact of Home Delivery (Matched Sample)**

	(1) Orders	(2) Log orders	(3) Log spending	(4) Log order value
D^{indv}	1.261***	0.108***	0.223***	0.021***
	(0.184)	(0.011)	(0.022)	(0.007)
$D^{indv} \times First$	2.587***	0.289***	0.642***	0.057***
	(0.297)	(0.016)	(0.029)	(0.007)
Customer FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Weighted	Yes	Yes	Yes	Yes
Adjusted R^2	0.404	0.580	0.525	0.442
Observations	387,624	387,624	339,171	308,270

Note: Standard errors are double-clustered at customer and time level. . *** p<0.01, ** p<0.05, * p<0.1

concern, we need to control for region-specific factors. In China, the city level is a fitting granularity, given the synchronized nature of local developments. Many government policies, designed in response to local social and economic situations, are both formulated and enforced at the city level. For example, when COVID-19 broke out in early 2020, Wuhan's lockdown was implemented citywide. Similarly, in the business realm, corporate strategies are frequently designed and executed with a city-centric focus. Therefore, we add city-month fixed effects into our baseline TWFE models to control for local trends.

$$y_{ict} = \alpha_i + \tau D_{it} + \gamma D_{it} \times First_{it} + \theta_{ct} + e_{it},$$

where c represents city. Table B2 shows that our results are robust.

Control for Local Community Trends One might still be concerned that the city level is not granular enough to capture all local factors. To further address this concern, we introduce more

Table B2 Impact of Home Delivery (Control for City Trends)

	(1) Orders	(2) Log orders	(3) Log spending	(4) Log order value
D^{indv}	1.340***	0.129***	0.270***	0.026***
	(0.237)	(0.009)	(0.019)	(0.005)
$D^{indv} \times First$	2.964***	0.272***	0.591***	0.059***
	(0.275)	(0.013)	(0.029)	(0.005)
Customer FE	Yes	Yes	Yes	Yes
City-Time FE	Yes	Yes	Yes	Yes
Adjusted R^2	0.481	0.613	0.549	0.447
Observations	2,398,896	2,398,896	2,099,034	1,819,032

Note: Standard errors are double-clustered at customer and time level. . *** p<0.01, ** p<0.05, * p<0.1

granular controls for local trends by incorporating station-month fixed effects:

$$y_{ist} = \alpha_i + \tau D_{it} + \gamma D_{it} \times First_{it} + \theta_{st} + e_{it},$$

where s represents station. Note that station-month fixed effects are very strong controls. With them being added, the model only exploits treatment status variations within the same Cainiao station. Therefore, stations without customer treatment status variations will be all discarded.[22] As a result, our sample size becomes much smaller. Nonetheless, despite the drastically different sample being used, Table B3 shows similar estimation results, which adds another layer of confidence to our empirical findings.

Table B3 Impact of Home Delivery (Control for Station Trends)

	(1) Orders	(2) Log orders	(3) Log spending	(4) Log order value
D^{indv}	1.654***	0.162***	0.339***	0.037***
	(0.261)	(0.012)	(0.026)	(0.007)
$D^{indv} \times First$	3.008***	0.267***	0.567***	0.049***
	(0.294)	(0.013)	(0.031)	(0.007)
Customer FE	Yes	Yes	Yes	Yes
Station-Time FE	Yes	Yes	Yes	Yes
Adjusted R^2	0.463	0.616	0.548	0.439
Observations	1,184,520	1,184,520	1,036,455	882,732

Note: Standard errors are double-clustered at customer and time level. . *** p<0.01, ** p<0.05, * p<0.1

[22] For example, stations that do not offer home delivery will be all dropped since everyone's treatment time path is a vector of zeros. Also, stations with only one customers being sampled into our data are also automatically removed by the TWFE model.

B.5. Discussion on COVID-19

One may concern about the potential impact of COVID on our findings as our sample period overlaps with the recent global pandemic. There are mainly two types of concerns: 1) Does COVID significantly alter people's behavior in such a way that our findings may lack generalizability? 2) Does COVID threaten our identification of causal effects?

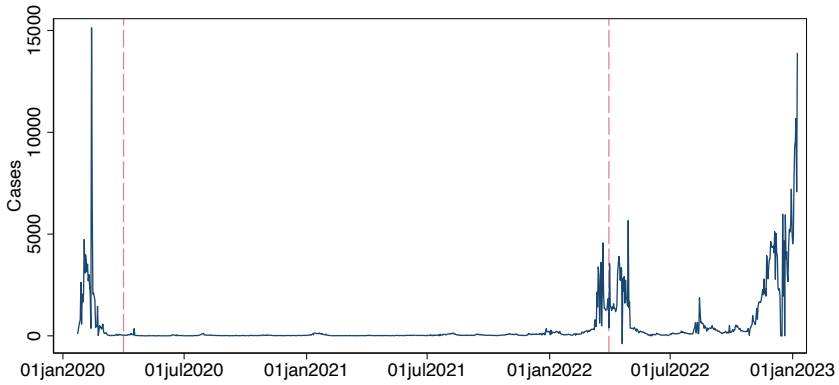


Figure B5 The Number of Newly Confirmed COVID Cases in Mainland China (Daily)

Source: JHU CSSE COVID-19 Data

First of all, it is important to note that COVID was effectively contained in China during the period we studied. Figure B5 displays the number of newly confirmed COVID cases in China from January 2020 to January 2023. Our sample period spans from April 2020 to March 2022, during which period there were very few new infections reported. Since most areas had zero cases for most of the time, COVID did not significantly impact the daily lives of most people. As a result, COVID should not pose a threat to the generalizability of our results.

Another potential concern is that COVID might simultaneously drive the adoption of home delivery and promote online shopping. However, given that COVID was well contained, this should not a significant issue. Moreover, for sporadic local outbreaks, most COVID related regulations were designed and enforced at the city level, while lockdowns.[23] were usually implemented at the community level. As a result, adding city-time or station-time fixed effects should sufficiently address these concerns, which which has already been examined in Section B.4.

Finally, we recognize that the Omicron wave hit Shanghai in March 2022, leading to a large-scale lockdown. Figure B5 confirms that the number of new cases quickly surged in March 2022. This Omicron shock in Shanghai may have intensified the regulations elsewhere, thereby significantly affecting people's lives nationwide. Despite this exceptional case, we are confident that primary

[23] In fact, lockdowns tend to disrupt the supply chain, which would lower the volume of online orders

findings remain valid, as excluding this period (i.e., March 2022) or Shanghai from our analysis does not alter our results.[24]

B.6. Customer Awareness and Spillover Effect

A potential concern for our identification strategy is the presence of spillover effects. Untreated customers at stations that adopted the home delivery program might have become aware that the service was being offered to others. This awareness could have negatively influenced their purchasing behavior—for example, by creating a sense of being "left out"—leading them to reduce their spending. If such an effect were present, our main analysis could overestimate the benefit of home delivery, as the comparison would be drawn against a control group that may have (partially) reduced its spending due to negative spillover.

To empirically assess this possibility, we conducted an additional placebo test. This test focuses exclusively on customers from the original control group and introduces a "fake" treatment based on whether their station had enrolled in the home delivery program. Specifically, we designated control-group customers at enrolled stations as the new "treatment" group, since they were potentially exposed to awareness of the program. Control-group customers at unenrolled stations, who were presumably unaware of the program, served as the new "control" group. This approach allows us to isolate spillover effects driven by customer awareness.

Specifically, we define the station-level entry indicator as follows:

$$D_{st}^{sta} = \begin{cases} 1, & \text{if station } s \text{ joined the program in or before period } t; \\ 0, & \text{otherwise.} \end{cases}$$

We then estimate the following TWFE linear regression model using the control group only, dropping all customers from the treatment group:

$$y_{ist} = \alpha_i + \tau D_{st} + \theta_t + e_{it}. \tag{8}$$

Table B4 reports the results. We find no statistically significant differences in spending patterns between these two groups. This null finding is consistent with the operational context of the program's rollout. During our study period, the service was in its early stages, with no widespread marketing or public announcements. The introduction of home delivery was managed as a station-level operational change, making it plausible that awareness among untreated customers remained very low. We therefore conclude that spillover effects are unlikely to be a significant source of bias in our main estimates.

[24] Results are available upon request.

Table B4 Impact of Station Entry on Untreated Customers

	(1) Orders	(2) Log orders	(3) Log spending	(4) Log order value
D^{sta}	0.023	-0.002	-0.012	-0.004
	(0.062)	(0.006)	(0.013)	(0.004)
Observations	1,604,856	1,604,856	1,404,249	1,175,188
Customer FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Adjusted R^2	0.586	0.586	0.520	0.428

Note: Standard errors are double-clustered at customer and time level. . *** p<0.01, ** p<0.05, * p<0.1

B.7. Alternative Identification Strategy

The home delivery program formally started in April, 2021. Since then, an increasing number of customers have been reached by the service. Nonetheless, we can use April, 2021 as the cutoff and assume all of the customers in the treatment group were immediately impacted. Under such an assumption, this scenario fits into the framework of the classic DiD where the timing of treatment is homogeneous. Therefore, we define

$$D_{it}^{unif} = \begin{cases} 1, & \text{if customer } i \text{ is in the treatment group and } t \geq 16 \text{ (April, 2021);} \\ 0, & \text{otherwise.} \end{cases}$$

We can then estimate the following TWFE linear regression model:

$$y_{it} = \alpha_i + \tau D_{it} + \theta_t + e_{it}. \tag{9}$$

Table B5 reports the estimation results, revealing that customers impacted by home delivery witnessed increases in their order volume, spending, and average order value. These effects are in line with our previous findings in Table 2 where the service impact is staggered, but are quantitatively smaller. Given our presumption that all customers were affected by the program immediately upon its launch, and considering many "late adopters" missed out on the early stage service, we likely underestimate the true impacts.

Parallel Trends Revisited. We can also plot the time trends for both the treatment and control groups. These trends are depicted in Figure B6, where the lines are mean-adjusted (we shift the control group line upward to align their pre-treatment means). We can see that even with the raw data, the time trends of the treatment and control groups are strikingly similar.

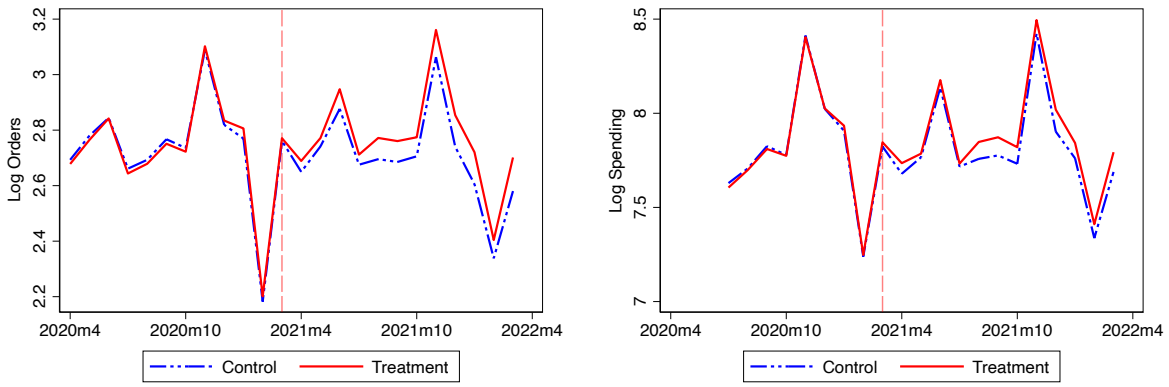
B.8. Weighted Regression

When performing customer-level analysis, linear regression models treat each customer as equally important a priori. This assumption may not always be desirable, especially when there is a strong belief that not all customers carry the same weight. For instance, in our sample, the bottom 25%

Table B5 Impact of Home Delivery on Customer-level Outcomes

	(1) Orders	(2) Log orders	(3) Log spending	(4) Log order value
D^{unif}	1.014*	0.084***	0.196***	0.033***
	(0.496)	(0.017)	(0.042)	(0.007)
Observations	2,400,000	2,400,000	2,100,000	1,819,983
Customer FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Adjusted R^2	0.467	0.604	0.541	0.446

Note: Standard errors are double-clustered at customer and time level. . *** p<0.01, ** p<0.05, * p<0.1



(a) Log Orders (b) Log Spending

Figure B6 Parallel Trends

of customers contribute to less than 1% of total sales. If the goal is to identify causal parameters that are more pertinent to aggregate outcomes, data variations from these low-frequency customers should be downgraded. To illustrate this point, suppose a high-frequency customer increases their spending by 17.9% and a low-frequency customer increases their spending by 41.2% (as indicated in Table 3). Assuming equal weights would result in an ATT of 29.5% ($=17.9\%\times 0.5+41.2\%\times 0.5$), a number quite similar to the estimates in Table 2 or Table B2. However, if the high-frequency customer contributes to 99% of the sales, the treatment effect on total spending should heavily lean towards that of high-frequency customers, which is expected to be around 18.1% ($=17.9\%\times 0.99+41.2\%\times 0.01$). To formalize this idea, we weight each customer by their pre-treatment order volume (i.e., the number of orders in fiscal year 2020) and then re-estimate Equation (1).

The results are reported in Table B6. Indeed, once weighted by order volume, the ATT estimates lean much more towards the ATT of high-frequency customers (as indicated in Table 3). This suggests that our findings remain robust, even when considering different weighting methods.

Table B6 Impact of Home Delivery (Weighted Regression)

	(1) Orders	(2) Log orders	(3) Log spending	(4) Log order value
D^{indv}	2.299***	0.112***	0.187***	0.018***
	(0.601)	(0.009)	(0.016)	(0.004)
$D^{indv} \times First$	4.119***	0.176***	0.298***	0.042***
	(0.540)	(0.012)	(0.022)	(0.006)
Customer FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Weighted	Yes	Yes	Yes	Yes
Adjusted R^2	0.532	0.614	0.580	0.536
Observations	2,400,000	2,400,000	2,100,000	1,819,983

Note: Standard errors are double-clustered at customer and time level. . *** p<0.01, ** p<0.05, * p<0.1

B.9. Alternative Variable Transformations

Log transformation has been widely used in empirical studies. Since many outcomes can take values in 0, using $\log(x)$ transformation is infeasible. Therefore, adding a small constant before taking the logarithm of data is a common practice. Following the convention, we use $\log(x+1)$ transformation in all previous analyses. However, when the values of outcome variables are small, log transformations can be sensitive to the choice of the constant. Therefore, we also explore alternative transformations including $\log(x+0.01)$ and $\text{arcsinh}(x)$. Table B7 shows that our findings are robust to alternative transformation methods.

Table B7 Impact of Home Delivery on Customer-level Outcomes

	Using $\log(x+0.01)$			Using $\text{arcsinh}(x)$		
	(1) Orders	(2) Spending	(3) Order value	(4) Orders	(5) Spending	(6) Order value
D^{indv}	0.277***	0.413***	0.029***	0.154***	0.291***	0.026***
	(0.020)	(0.031)	(0.005)	(0.011)	(0.021)	(0.005)
$D^{indv} \times First$	0.597***	0.915***	0.067***	0.331***	0.653***	0.063***
	(0.036)	(0.056)	(0.005)	(0.017)	(0.037)	(0.005)
Observations	2,400,000	2,100,000	1,819,983	2,400,000	2,100,000	1,819,983
Customer FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.470	0.461	0.424	0.597	0.531	0.446

Note: Standard errors are double-clustered at customer and time level. . *** p<0.01, ** p<0.05, * p<0.1

B.10. Analysis Without Log Transformation

In our main analyses, we only use log-transformed outcome variables for spending and average order value, which is standard practice in empirical research. This transformation enhances inter-

pretability, allowing coefficients to approximate percentage changes (semi-elasticities) that are more intuitive and more transferrable across business contexts. It also addresses the right-skewness commonly found in transactional data by normalizing the distribution and reducing the impact of extreme values, thereby improving estimation efficiency and stability.

To examine the results without log transformation, we estimate Equation (1) using the raw (unlogged) outcomes. However, regressions on such skewed data often produce unstable estimates with large standard errors. To mitigate this issue, we winsorize the outcome variables at the top 0.1%, 1%, and 2% within each customer group and time period, and see how the estimation evolves.

Table B8 presents the results. Without winsorization, we found positive effects but with extremely large standard errors. In contrast, the estimates become more stable and significant after winsorization, especially at the 1% and 2% thresholds.

Table B8 Impact of Home Delivery on Spending and Order Value

	Monthly Spending				Average Order Value			
	(1) Raw	(2) Top 0.1%	(3) Top 1%	(4) Top 2%	(5) Raw	(6) Top 0.1%	(7) Top 1%	(8) Top 2%
D^{indv}	306.426	184.302**	162.682*	156.407*	0.636	0.064	1.124	1.379**
	(380.382)	(75.468)	(79.699)	(82.163)	(10.811)	(0.937)	(0.719)	(0.640)
$D^{indv} \times First$	575.653***	516.507***	478.907***	444.957***	2.615	0.249	2.193***	2.598***
	(115.828)	(75.806)	(67.426)	(63.598)	(5.329)	(0.965)	(0.615)	(0.582)
Observations	2,100,000	2,100,000	2,100,000	2,100,000	1,819,983	1,819,983	1,819,983	1,819,983
Customer FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.00206	0.463	0.518	0.540	-0.00748	0.221	0.310	0.346

Note: Standard errors are double-clustered at customer and time level. . *** p<0.01, ** p<0.05, * p<0.1

C. Causal Forest

This appendix complements Section 6.3 by briefly introducing causal forests. Besides DML, the causal forest approach is another popular choice for estimating CATE functions.[25] Like DML, causal forests offer notable advantages over conventional econometric methods; in particular, the flexibility—enabled by tree-based structure—to include features in a nonparametric way.

[25] Meta-learners are another important category of methods for CATE estimation. We experimented with meta-learner methods such as the X-learner (Künzel et al. 2019), the R-learner (Nie and Wager 2021), and the DR-learner (Foster and Syrgkanis 2023). Their performance sits between that of CF and of DML. Since an exhaustive model comparison is beyond our scope, additional results are available upon request.

The building blocks of causal forests are causal trees, which are analogous to regression trees but adopt a different split criterion. For a given tree structure, a causal tree estimates the CATE as follows:

$$\hat{\tau}(x) = \frac{1}{|\{i : D_i = 1, \mathbf{X}_i \in L(x)\}|} \sum_{\{i:D_i=1,\mathbf{X}_i\in L(x)\}} Y_i - \frac{1}{|\{i : D_i = 0, \mathbf{X}_i \in L(x)\}|} \sum_{\{i:D_i=0,\mathbf{X}_i\in L(x)\}} Y_i, \quad (10)$$

where $L(x)$ is the leaf that contains x. Essentially, it computes the difference in means between treatment and control groups within a leaf.[26] When splitting a parent node, the causal tree aims to maximize the difference in treatment effect between two child nodes.

Similar to the process of aggregating regression trees to form random forests, causal trees can also be used to grow causal forests. Specifically, suppose a causal forest consists of B causal trees, with each tree gives an estimate $\hat{\tau}_b(x)$. Then the causal forest predicts the CATE by taking the average: $\hat{\tau}(x) = B^{-1} \sum_{b=1}^{B} \hat{\tau}_b(x)$. This ensemble approach capitalizes on the individual strengths of causal trees, particularly in identifying heterogeneous treatment effects, while simultaneously reducing the risk of overfitting and variance that might be present in a single causal tree. Most importantly, the causal forest estimator of CATE has been shown to be point-wise consistent and asymptotically normal (Wager and Athey 2018).[27] We implement the causal forest as described in Athey and Wager (2019), a version that is more tailored to observational data.

To further control for time-invariant confounders, we adopt the first difference causal forests approach (Wang 2022), taking the difference between purchases after and before receiving home delivery service for each customer and using it as the outcome variable in causal forests. We include the estimation results of both the causal forests (CF) and the first-difference causal forests (FD-CF) for comparison.

D. Additional Results from Causal Machine Learning

In this section, we presents additional results from causal machine learning models. We first show how we can estimate traditional causal inference metrics such as the ATT, ATU, and ATE in Section D.1. While the ATT offers valuable insights for program evaluation, the ATU can inform future investment opportunities. Specifically, we illustrate how to use ATU to guide future investment decisions in Section D.2.

[26] In contrast, a regression tree typically predicts the outcome variable by averaging all observations in the leaf: $\hat{y}(x) = \sum_{\{i:\mathbf{X}_i\in L\}} Y_i / |\{i : \mathbf{X}_i \in L\}|$.

[27] To establish consistency and normality, it also requires the trees to be honest, which means a sample can be used either to grow a tree or to estimate τ, but not for both purposes. A similar sample-splitting procedure, known as cross-fitting, is adopted by DML.

D.1. Estimating ATT, ATU, and ATE

Causal machine learning methods estimate CATE, a much more granular object than ATE. Thanks to the flexibility of CATE functions, we can easily aggregate CATE to gain aggregate effects such as the ATT, ATU, and ATE, and perform statistical inference. Table D1 reports the results.

In particular, the estimated ATTs are comparable to the treatment effects produced from DiD, since DiD also identifies ATT. In Table 2, the coefficient of D^{indv} on log orders is 0.13, which implies an 13.88% increase in order volume.[28] Although we use different periods of data, the estimates are of similar magnitude.

Table D1 Estimation Results

	Treated population			Untreated population			Full sample		
Model	ATT	$\mathbb{E}Y(0\|1)$	ATT (%)	ATU	$\mathbb{E}Y(0\|0)$	ATU (%)	ATE	$\mathbb{E}Y(0)$	ATE (%)
DML	7.07	53.79	13.14%	5.42	33.19	16.34%	5.74	37.12	15.45%
FD-DML	7.36	53.50	13.76%	5.54	33.19	16.69%	5.89	37.06	15.88%
NP-DML	7.60	53.26	14.27%	5.63	33.19	16.96%	6.01	37.02	16.22%
CF	8.38	52.47	15.98%	6.06	33.19	18.27%	6.51	36.87	17.65%
FD-CF	7.91	52.95	14.94%	5.95	33.19	17.93%	6.32	36.96	17.11%
DiD			13.88%						

D.2. Regional Heterogeneity and Expansion Strategy

If the retailer decides to expand the coverage of home delivery service, where should they start? From an investment perspective, money and resources should be invested in areas where yield the highest return on investment (ROI). In the language of causal inference, if more units of treatment were available, they should be assigned to those with higher ATU.[29]

In China, many business strategies operates at the level of city. Therefore, we estimate the ATU at the city level. Cities with higher ATU would experience a higher sales growth if treatments were assigned. Using the FD-DML model, Figure D1 displays the city-level ATU and its confidence interval for top 5 and bottom 5 cities, where we can observe significant heterogeneity across cities. Table D2 further illustrates the magnitude in terms of percentage change. Clearly, the services in those bottom cities are already saturated in terms of ROI; providing additional coverage yields minimal marginal returns. On the other hand, cities like Taizhou demonstrate strong growth potential. Therefore, if Cainiao decides to expand the program, these top cities should be prioritized.

D.3. User Features and Importance Analysis

We list the set of user features in our causal machine learning models in Table D3.

[28] Approximated by $e^{0.13} - 1$.

[29] For simplicity, we assume the cost of treatment is uniform. A more careful analysis should be based on the capacity-aware uplift model, which fully considers the heterogeneous cost structure.

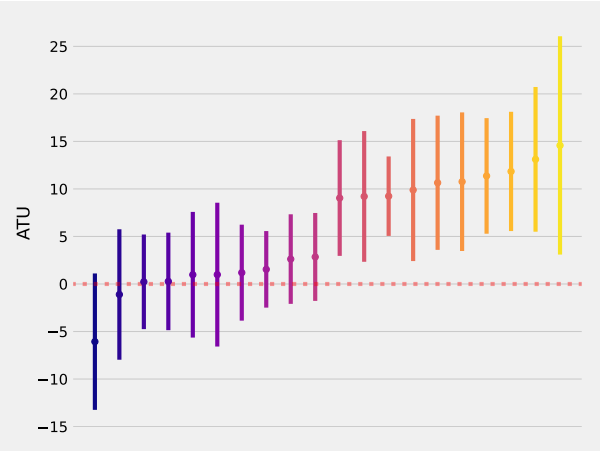


Figure D1 **ATU by Cities: Top 5 and Bottom 5**

Table D2 **Top 5 Cities with Highest ATU**

	ATU	CI(95%)	$\overline{Y}$	ATU (%)
Taizhou	14.58	(3.10, 26.06)	37.66	38.71
Changchun	13.11	(5.50, 20.73)	30.75	42.64
Tianjin	11.84	(5.57, 18.12)	40.20	29.46
Nanning	11.37	(5.29, 17.45)	28.99	39.22
Fuzhou	10.76	(3.47, 18.05)	42.14	25.54

Table D3 **User Features Used in Causal Machine Learning Models**

Variable	Description / Values	Mean	Std. Dev.
Continuous			
Number of orders	2020Q4, scaled	1	1.407
Log spending	2020Q4, scaled	1	0.311
Primary city share	Share of packages delivered to the customer's primary city of residence (2020Q4)	0.966	0.058
Categorical			
Gender	Female, Male	—	—
Age	1–18, 19–25, 26–30, 31–35, 36–40, 41–50, 51–60, ≥61	—	—
Education	Associate Degree, High School, Doctoral Degree, Bachelor's Degree, Master's Degree	—	—
Occupation	Self-employed, Civil Servant, Company Employee, Medical Worker, Media Worker, Student, Worker, Teacher/Faculty, Researcher, Financial Worker	—	—
City	Top 50 cities retained; remaining cities grouped into "Other" category	—	—

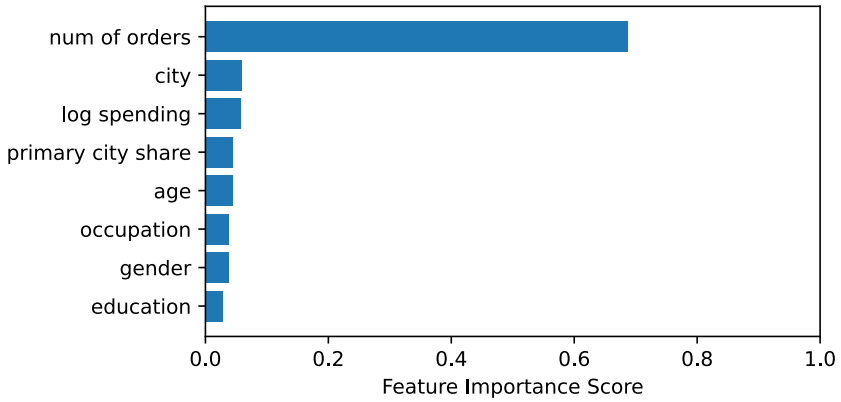


Figure D2 **Feature Importance for Outcome Y**

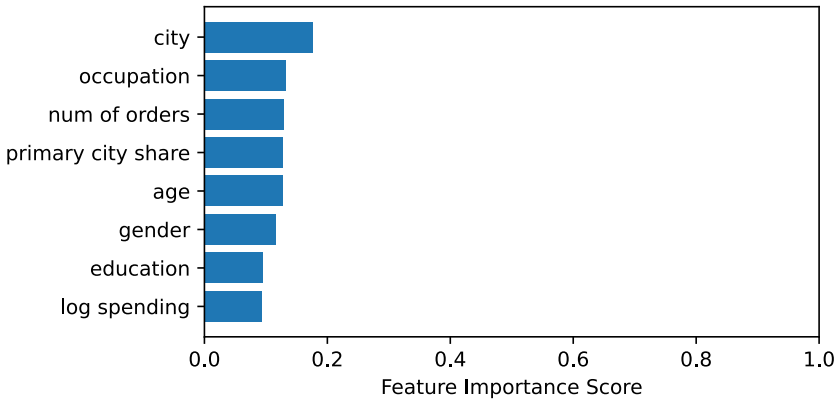


Figure D3 Feature Importance for CATE

To further understand the role of these features, we evaluate their importance in two key tasks: (1) predicting the outcome variable Y (number of orders in 2021Q4) and (2) contributing to the second-stage residualized regression for CATE estimation. For each task, we fit an XGBoost regression model and compute feature importance scores based on the average gain from splits across all trees.[30]

We visualize the resulting feature importance scores. Figure D2 shows that the number of orders in the previous period (2020Q4) is the dominant predictor of the number of orders in 2021Q4, which is unsurprising given that past purchasing behavior is typically the strongest predictor of future. In contrast, Figure D3 highlights a more balanced set of drivers contributing to CATE estimation.

The CATE function projects true individual treatment effects onto the space spanned by observed features. Expanding this feature space with more informative variables would enhance prediction accuracy. For example, customer-to-station distance is a likely moderator of treatment effectiveness, since the inconvenience of self-pickup increases with distance. Incorporating such variables therefore presents a valuable feature-engineering opportunity, improving CATE estimation and, in turn, refining targeting policies in practice.

E. Nuisance Parameter Estimation in DML

Accurate estimation of nuisance parameters—namely, the outcome model $\hat{Y}(X)$ and the treatment model $\hat{D}(X)$—is critical for the performance of DML. These nuisance components are used to orthogonalize the treatment effect estimation, ensuring that the final causal estimates are robust to model misspecification and high-dimensional confounding. The choice of predictive models for these components therefore plays a central role in the quality and reliability of the causal inference.

[30] Feature importance scores from XGBoost reflect the average improvement in model performance (gain) contributed by each feature when used in decision tree splits.

However, it is important to emphasize that evaluating the performance of these nuisance models is only an intermediate step—it does not, by itself, validate the accuracy of the resulting CATE estimates.

E.1. Model Classes Considered

We consider a representative set of industry-standard predictive models:

- *Linear models*: Including standard linear regression and its regularized variant, LASSO. To increase model expressiveness, we also evaluate an extended specification that includes all second-order terms of the features, capturing both squared terms and pairwise interactions.
- *Tree-based models*: These range from simple decision trees to more advanced ensemble methods such as random forests and gradient-boosted trees (XGBoost). For XGBoost, we also perform hyperparameter tuning using a grid search over the number of estimators, maximum tree depth, and learning rate.
- *Neural networks*: We experiment with three feedforward architectures of increasing complexity, each using ReLU activation functions. The small network includes two hidden layers with 32 and 16 neurons $(32, 16)$; the medium network uses $(64, 32)$; and the large network includes three hidden layers with $(128, 64, 32)$. These configurations allow us to explore the trade-offs between model capacity, overfitting, and generalization.

E.2. Model Evaluation

All models are trained to minimize mean squared error (MSE) on the training set. For models requiring hyperparameter tuning, we further split the training set into a training subset and a validation subset, selecting hyperparameters that minimize validation MSE. Models are then evaluated on a common holdout test set to ensure comparability. Performance is assessed using two complementary metrics: root mean squared error (RMSE) and mean absolute error (MAE).

We report evaluation results for the outcome model Y in Figure E1 and the treatment model D in Figure E2.[31] Both figures show that modern machine learning methods significantly outperform traditional linear models. In particular, the tuned XGBoost model achieves the lowest RMSE among all models. Tree-based ensembles and small neural networks also perform well, while large neural networks and unregularized linear models with interactions tend to overfit, leading to poor generalization. These results highlight the importance of model selection and regularization when estimating nuisance parameters for DML.

[31] For the FD-DML, the nuisance parameter estimation involves predicting ΔY. Evaluation metrics for ΔY are very similar to that of Y and are available upon request.

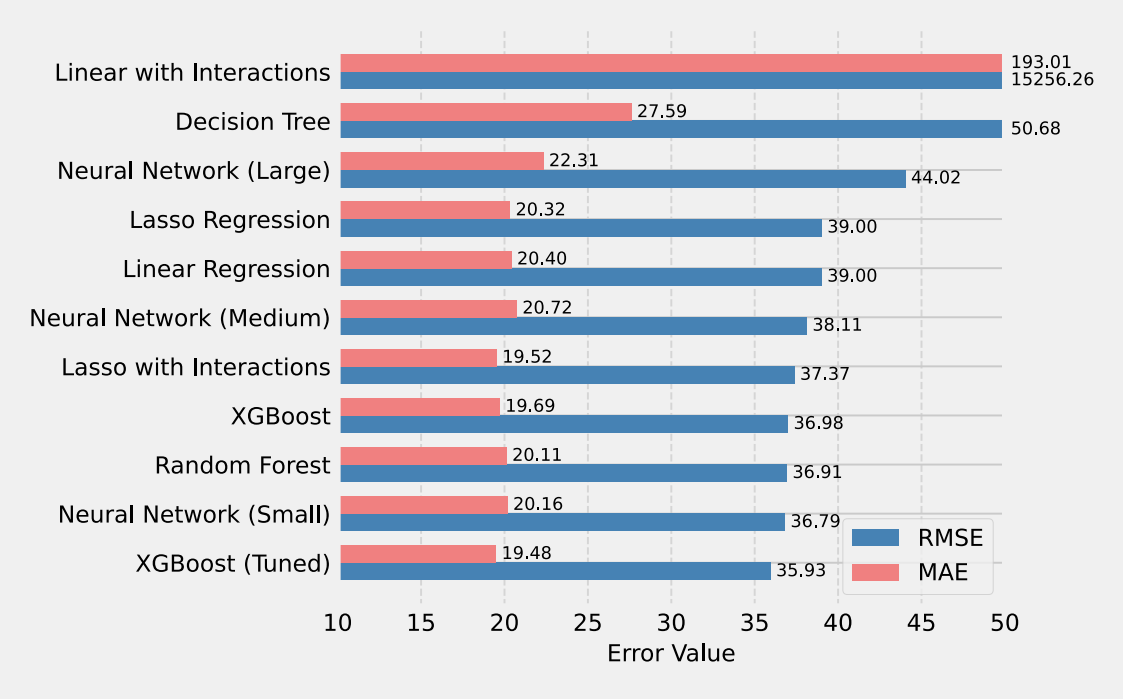


Figure E1 Predictive Accuracy for Outcome Y

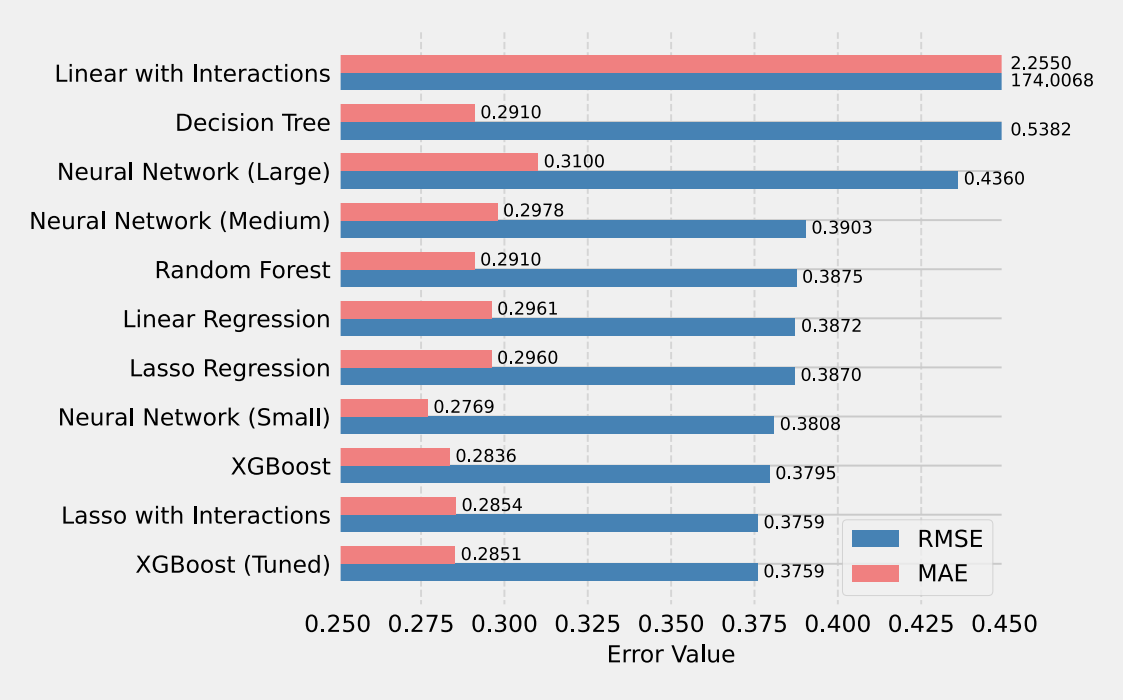


Figure E2 Predictive Accuracy for Treatment D

E.3. Performance Insights

Our evaluation of predictive models yields three key findings:

1. **Linear models underperform**: Standard linear regression performs relatively poorly, as it cannot capture the complex, nonlinear dependencies in the data.

2. **Interactions in linear models require caution**: Adding second-order terms can introduce flexibility but also increase the risk of multicollinearity and overfitting—especially without regularization. We find that linear regression with unpenalized interactions can lead to extremely unstable and inaccurate predictions, reinforcing the need for more powerful machine learning methods.

3. **Machine learning models perform well when properly configured**:

• LASSO performs well when paired with thoughtful feature engineering, such as adding interaction terms. Without interaction terms, however, it offers limited—if any—improvement over standard linear regression in our context. This is likely because the initial feature set has already been manually curated to include only informative variables, based on our domain knowledge. As a result, the marginal benefit of LASSO's automatic variable selection is modest.

• Tree-based models also perform well. While simple decision trees underperform due to their high variance, ensemble methods such as random forests and XGBoost significantly improve predictive accuracy by reducing variance and/or bias through model aggregation. In particular, a well-tuned XGBoost consistently achieves top performance in terms of both RMSE and MAE.

• Neural networks also exhibit competitive performance. However, larger architectures tend to overfit, resulting in reduced accuracy. This observation is consistent with prevailing practice: for traditional tabular data, tree-based ensemble methods often outperform neural networks, which are more sensitive to overfitting and typically require larger datasets to generalize effectively. While additional regularization techniques—such as dropout or weight decay—may help mitigate overfitting, they require careful tuning and increase the overall complexity of model development.

E.4. Summary

These results highlight the limitations of traditional linear models and the value of modern machine learning techniques in capturing the complex, nonlinear relationships often present in real-world data. Although the DML framework is theoretically robust to moderate inaccuracies in nuisance parameter estimation, it remains critical in practice to ensure that the models used for this purpose are well-specified and performant. Higher predictive accuracy in estimating nuisance components can lead to more reliable and efficient causal effect estimation, as it supports the validity of the orthogonalized score functions central to DML.

The choice of machine learning model for nuisance parameter estimation should ultimately depend on the characteristics of the data and the structure of the problem at hand. Based on our empirical comparison, we select the best-performing XGBoost model to estimate nuisance components consistently across all DML implementations in Section 6.

F. Evaluating Routing-Aware Uplift Model

To empirically evaluate the trade-offs discussed in Section 7.3, we compare multiple customer selection strategies under a fixed delivery distance constraint. Each strategy aligns with a quadrant of the 2×2 framework introduced earlier, defined by dimensions of *value-awareness* and *routing-awareness*.

F.1. Experimental Setup

We conduct our evaluation in the West Lake District of Hangzhou using real customer data. Specifically, we randomly sample twenty customer addresses from actual delivery records to capture representative spatial patterns across residential and commercial areas. We focus on a small, controlled instance for two reasons. First, it enables exact route optimization via dynamic programming (DP), allowing us to obtain a globally optimal solution[32] that serves as an unambiguous benchmark for comparing our value-aware routing against alternative approaches. Second, it facilitates clear visualization and interpretation of routing outcomes and trade-offs.

Each customer is assigned a predicted Individual Treatment Effect (ITE), measuring the expected operational benefit derived from serving that customer. These ITEs are drawn from a Gaussian distribution with a mean normalized to 1, and a calibrated standard deviation reflecting the heterogeneity observed in real-world ITE estimates from FD-DML. It is important to note that some customers have negative ITEs.

A Cainiao Station is located at the geographic center of the service area. The courier starts and ends its route at the station and must not exceed a total travel distance of 5 kilometers. Pairwise distances between all points are computed using geodesic formulas to approximate travel distances in kilometers based on actual latitude and longitude coordinates.[33] Figure F1 depicts the spatial layout, where each circle represents a customer, and the circle size corresponds to the magnitude of their assigned ITE.

F.2. Implementation Details

We implement and evaluate six distinct routing strategies, each differing in their approach to integrating ITEs and routing efficiency:

- Value-Aware Routing (Max Uplift DP): A dynamic programming (DP) approach that jointly optimizes customer selection and routing to maximize the cumulative ITE within the distance constraint. This strategy identifies the globally optimal route among the 20-customer set.

[32] In large-scale practical settings, DP may become computationally intractable, which would necessitate the use of scalable approximation methods. Nonetheless, the core insights from our analysis are expected to hold.

[33] We also experimented with road-based distances from a third-party routing API on a smaller instance, which produced similar results.

• Value-Unaware Routing (Max Coverage DP): Another DP-based algorithm that prioritizes serving the maximum number of customers, irrespective of their ITE values, subject to the same distance constraint.

• Value-Based Ranking (Highest ITE First): A heuristic prioritizing customers solely based on descending ITE values. Customers are sequentially added to the route as long as the distance constraint is not violated. This method is value-aware but routing-unaware, as geographic proximity is ignored.

• Random Selection: Customers are selected in a random order to be included in the route, provided that the inclusion of each customer and completion of the tour does not exceed the maximum route distance. This value- and routing-unaware strategy serves as a weak baseline.

Additionally, we implement two further heuristics:

• Distance-Based Ranking (Nearest First): This heuristic selects customers strictly based on their proximity (e.g., shortest distance from the current endpoint of the route or from the station). Customers are added sequentially if the distance constraint is maintained. This strategy is value-unaware, ignoring customer ITE values, but partially routing-aware.

• Nearest High-ITE: This hybrid heuristic first filters customers to include only those in the top 50% based on their ITE values. It then applies a nearest-neighbor routing logic among this subset to construct a route within the distance limit. This hybrid approach is partially value- and routing-aware.

All implemented strategies generate a complete tour, starting and ending at the station. The output for each policy includes the total ITE accrued from the served customers, the total distance traveled, the number of customers served, and the derived efficiency metric of ITE per kilometer (ITE/km). For the Random Selection strategy, results reflect averages computed over multiple simulation runs.

Table F1 Performance Comparison of Routing Strategies

Policy	Customers	Distance	Total ITE	ITE/km
Value-Aware Strategies				
Value-Aware Routing	14/20	4.94/5	23.40/20	4.74
Nearest High-ITE	10/20	4.91/5	21.39/20	4.35
Value-Based Ranking	6/20	4.87/5	13.95/20	2.87
Value-Unaware Strategies				
Value-Unaware Routing	16/20	4.17/5	15.37/20	3.69
Distance-Based Ranking	9/20	3.76/5	6.42/20	1.71
Random	7.8/20	4.87/5	6.03/20	1.23

F.3. Quantitative Results

Table F1 summarizes the performance metrics for each strategy. Optimization-driven methods, particularly Value-Aware Routing (Uplift DP), consistently outperform simpler heuristics and benchmarks, achieving higher total ITE and better ITE/km efficiency.

Figure F1 visualizes the routing paths generated by Value-Aware Routing and Value-Unware Routing, revealing notable differences. Both strategies favor customer clusters and avoid two isolated customers (marked as orange circles) located at the corners, as these are routing-inefficient. However, Value-Unaware Routing selects two nearby customers with negative ITEs while skipping two high-ITE (green) customers that are farther and more isolated. In contrast, Value-Aware Routing prioritizes the high-value customers despite their distance, resulting in a more effective allocation.

This comparison highlights the fundamental distinction between the two approaches. While Value-Unaware Routing may serve more customers overall, it captures less cumulative ITE. These results highlight the importance of causal machine learning in driving value-based treatment decisions in the context of routing problems.

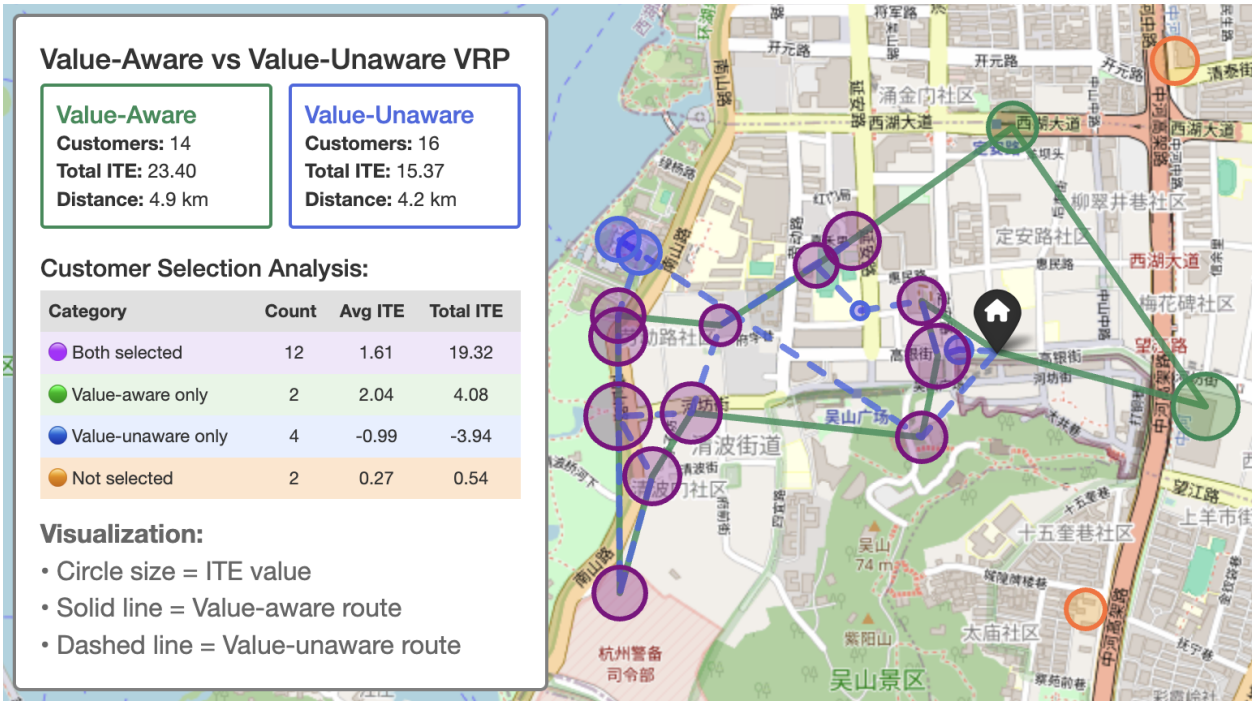
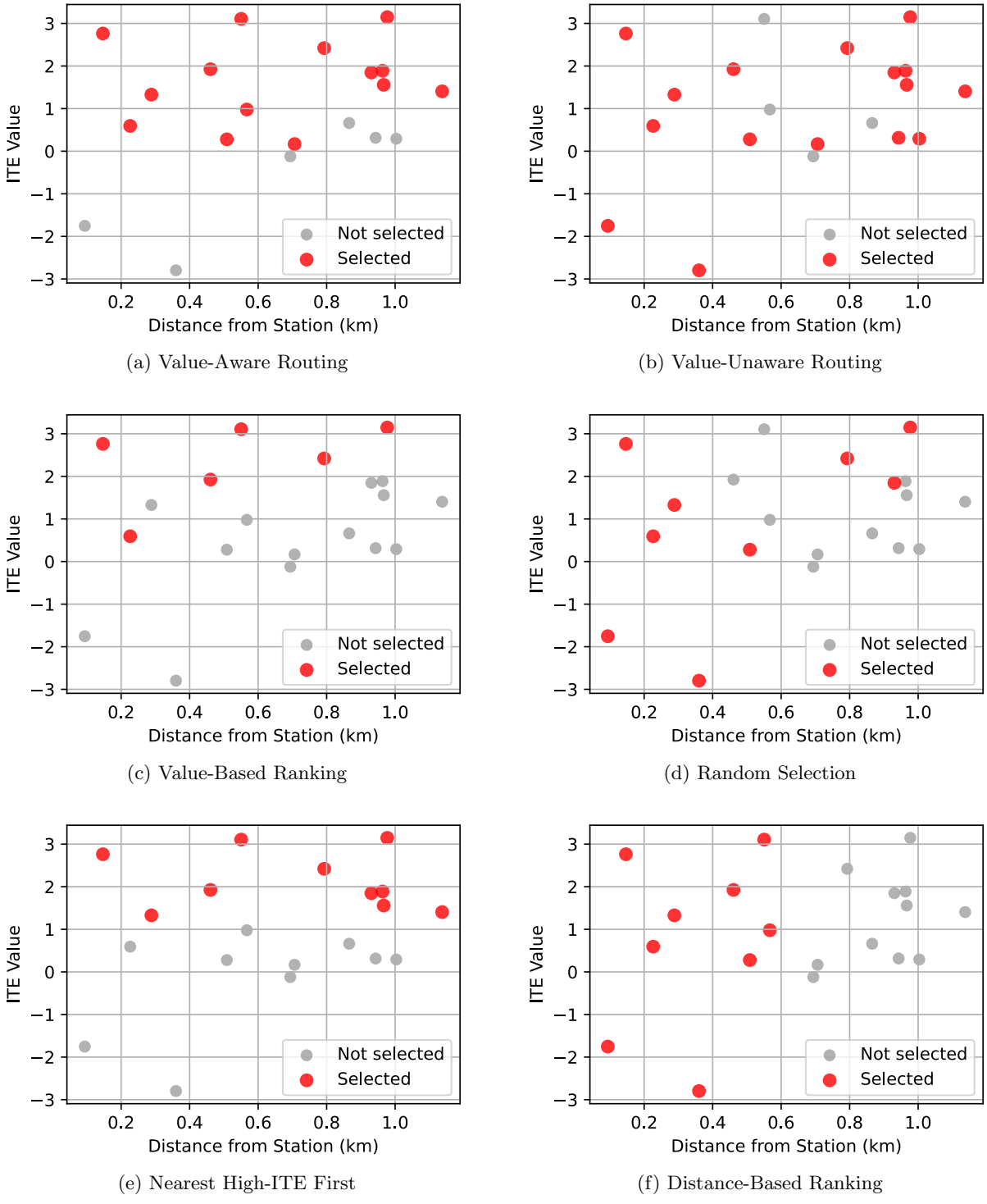


Figure F1 **Value-Aware vs. Value-Unaware Optimal Routing**

Figure F2 presents scatter plots that compare selected versus unselected customers in the ITE–distance space across different strategies. The visualizations reveal that value-aware strategies effectively identify and include high-ITE customers even when they are located farther from the

station. In contrast, routing-only strategies tend to favor customers situated closer to the station, regardless of their ITE, often overlooking more valuable but distant customers.



(a) Value-Aware Routing

(b) Value-Unaware Routing

(c) Value-Based Ranking

(d) Random Selection

(e) Nearest High-ITE First

(f) Distance-Based Ranking

Figure F2 ITE vs. Distance Across Customer Selection Strategies

F.4. Summary

Our findings highlight the critical role of causal-based customer response in improving operational efficiency through value-aware decision-making. Value-aware strategies consistently identify and prioritize high-impact customers—even those located farther from the station—whereas routing-focused approaches tend to favor proximity, often at the expense of overall treatment effectiveness. These results demonstrate the importance of integrating treatment effect heterogeneity into routing decisions, particularly in resource-constrained logistics settings. Realizing the full potential of causal machine learning, however, requires the incorporation of domain expertise and operational constraints, just as exemplified in the design of capacity-aware uplift models.

Appendix References

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, Stefan Wager. 2021. Synthetic difference-in-differences. *American Economic Review* **111**(12) 4088–4118.

Athey, Susan, Stefan Wager. 2019. Estimating treatment effects with causal forests: An application. *Observational studies* **5**(2) 37–51.

Callaway, Brantly, Pedro H C Sant'Anna. 2021. Difference-in-Differences with multiple time periods. *Journal of Econometrics* **225**(2) 200–230.

Cui, Ruomeng, Meng Li, Qiang Li. 2020. Value of high-quality logistics: Evidence from a clash between sf express and alibaba. *Management Science* **66**(9) 3879–3902.

Fisher, Marshall L, Santiago Gallino, Joseph Jiaqi Xu. 2019. The value of rapid delivery in omnichannel retailing. *Journal of Marketing Research* **56**(5) 732–748.

Foster, Dylan J, Vasilis Syrgkanis. 2023. Orthogonal statistical learning. *The Annals of Statistics* **51**(3) 879–908.

Künzel, Sören R, Jasjeet S Sekhon, Peter J Bickel, Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences* **116**(10) 4156–4165.

Nie, Xinkun, Stefan Wager. 2021. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108**(2) 299–319.

Sun, Liyang, Sarah Abraham. 2021. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics* **225**(2) 175–199.

Wager, Stefan, Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* **113**(523) 1228–1242.

Wang, Guihua. 2022. The effect of medicaid expansion on wait time in the emergency department. *Management Science* **68**(9) 6648–6665.