The Economic Consequences of Data Privacy Regulation: Empirical Evidence from GDPR*

Guy Aridor† Yeon-Koo Che‡ Tobias Salz§

April 6, 2020

Abstract

This paper studies the effects of the EU's General Data Protection Regulation (GDPR) on the ability of firms to collect consumer data, identify consumers over time, accrue revenue via online advertising, and predict their behavior. Utilizing a novel dataset by an intermediary that spans much of the online travel industry, we perform a difference-in-differences analysis that exploits the geographic reach of GDPR. We find a 12.5% drop in the intermediary-observed consumers as a result of the new opt-in requirement of GDPR. At the same time, the remaining consumers are observable for a longer period of time. We provide evidence that this pattern is consistent with the hypothesis that privacy-conscious consumers substitute away from less efficient privacy protection (e.g, cookie deletion) to explicit opt out, a process that would reduce the number of artificially short consumer histories. Further in keeping with this hypothesis, we observe that the average value of the remaining consumers to advertisers has increased, offsetting most of the losses from consumers that opt out. Finally, we find that the ability to predict consumer behavior by the intermediary's proprietary machine learning algorithm does not significantly worsen as a result of the changes induced by GDPR. Our results highlight the externalities that consumer privacy decisions have both on other consumers and for firms.

Keywords: GDPR, Data Privacy Regulation, E-Commerce
JEL Codes: L50; K20; L81.

*We would like to thank Daron Acemoglu, Francesco Decarolis, Glenn Ellison, Sara Ellison, and seminar participants at Columbia for helpful comments. We would further like to thank William Nelson for his help. Krista Moody provided outstanding research assistance. All errors are our own.

†Columbia University, Department of Economics. Email: g.aridor@columbia.edu
‡Columbia University, Department of Economics. Email: yeonkooche@gmail.com
§MIT, Department of Economics. Email: tsalz@mit.edu

1 Introduction

Technological advances in the past several decades have led to enormous growth in the scale and precision of consumer data that firms collect. These advances have been followed by progress in machine learning and other data processing technologies that have allowed firms to turn data into successful products and services and earn vast economic returns along the way.[1] However, at the same time, there has been an increasing number of high profile data breaches and a growing feeling of despondency amongst consumers who lack control over this process.[2,3] Against this backdrop, government regulators have proposed and enacted data privacy regulation that empowers consumers to have more control over the data that they generate. The European Union was the first to enact such legislation - the General Data Protection Regulation - which has served as a blueprint for privacy legislation in California, Vermont, Brazil, India, Chile, and New Zealand.[4] However, we lack empirical evidence on the effectiveness and broader impact of such regulation. Such evidence is critical not only for guiding the design of upcoming regulation, but also to understand fundamental questions in the economics of privacy.

This paper empirically studies the effects of the EU's General Data Protection Regulation (GDPR), in particular, its requirement that consumers be allowed to make an informed, specific, and unambiguous consent to the processing of their data. The *consent requirement* constitutes the front-line of privacy protection for consumers and potentially threatens the data-driven business model of firms. The consent option provides a simple but effective means of protecting privacy: by denying consent, a consumer can block a website from collecting personal data and sharing it with third-party affiliates. At the same time, consent denial inhibits firms from tracking consumers across time and across websites, thereby building historical profiles of consumers. Hence, consumers exercising their consent right can significantly hamper these firms' ability to learn and predict consumer behavior and target their services and advertising accordingly.

Our investigation focuses on three broad questions. First, *to what extent do consumers exercise*

[1] Several popular press outlets have gone as far as stating that "data is the new oil" meaning that the world's most valuable resource is now data, not oil (e.g. The world's most valuable resource is no longer oil, but data `https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data`. Retrieved on January 9th, 2020.).

[2] There have been many but among the most prominent are the Cambridge Analytica and Equifax data breaches. Cambridge Analytica harvested the personal data of millions of people's Facebook profiles without their consent and used it for political advertising purposes. The Equifax data breach exposed the names, dates of birth, and social security numbers of 147 million individuals.

[3] We Hate Data Collection. That Doesn't Mean We Can Stop it. `https://www.nytimes.com/2019/11/15/opinion/privacy-facebook-pew-survey.html`. Retrieved on January 3rd, 2020.

[4] While such regulation is not entirely novel, the scope and robustness of previous regulation pales in comparison to that of GDPR. Several states in the United States and countries around the world are debating and implementing their own privacy regulations with similar scope and stipulations as GDPR. For more information on the specifics of the various laws and how they relate to GDPR: 6 New Privacy Laws Around The Globe You Should Pay Attention To. `https://piwik.pro/blog/privacy-laws-around-globe/`. Retrieved on March 10th, 2020.

the consent right enabled by GDPR? Anecdotal and survey evidence suggests that consumers value their privacy, but are they willing to take action to protect it when presented with a simple effective means? Or, do they simply ignore the option and give away their personal data even at little cost?[5] We do not yet have clear empirical answers to these questions.

Second, *how does GDPR change the composition of consumers observed by firms?* Even prior to GDPR, consumers were able to protect their privacy by utilizing browser-based privacy protection means. However, utilizing these privacy means would result in the underlying data still being sent to the website, but it would be associated with different identifiers so that the website could not link this data to the same consumer. These "spurious" consumer footprints are difficult to distinguish from genuine consumer footprints left by consumers who do not adopt these privacy means. This process creates noise in the data observed by firms that could make it difficult for them to track consumers and predict their behavior. Under the GDPR regime, however, the same consumers may simply opt out, in which case they do not leave any footprints, and this could in principle make the remaining consumers more easily trackable and identifiable. This raises an interesting question of externalities created by privacy tools on the other consumers and for the firms. To the best of our knowledge, these forms of *privacy externalities* not only differ from those recognized in the theoretical literature (Choi, Jeon and Kim, 2019; Acemoglu et al., 2019; Bergemann, Bonatti and Gan, 2019) but more importantly have never been empirically identified.

Third, *how does the GDPR privacy protection impact firms that rely crucially on consumer data?* Specifically, how does consumer opt out affect firms' abilities to learn and predict consumer behavior and to provide targeted advertising? And how do advertisers react to such a change? Do they bid more or less for reaching consumers made available for them as a result? Evidently, opt outs will reduce the *scale* of data available to firms which may lead to firms' prediction capabilities to suffer as result. But at the same time, the possible change in the composition of consumers could presumably change the *quality* of data. If the remaining consumers are more easily tracked and attributed to purchase, this could possibly increase the value of consumers to advertisers and thus can make up for the possible loss in scale.

To answer these questions, we use the data provided by an anonymous intermediary that operates in more than 40 countries and contracts with many of the largest online travel agencies and travel meta-search engines. The dataset is uniquely suited for the current inquiries in several respects. An integral part of the intermediary's business is to predict the likelihood of purchase of each consumer upon every visit, based on the identifiable history of past behavior, which it uses to personalize the experience of consumers on the website. The data links consumers' behavior

[5] A prevalent theme in the economics of privacy literature consistently finds a privacy paradox - the apparent inconsistency between individual's strong stated preferences for privacy and their willingness to give away personal information at little cost (Acquisti, Taylor and Wagman, 2016). This directly implies that a natural hypothesis is that consumers may ask legislators for such privacy means but, ultimately, make little use of them.

across time and across websites using cookies (set by the intermediary), small files stored attached to a consumer's web browser that allow the intermediary to identify consumers. We observe (in anonymized and aggregated form) the same rich consumer information as the intermediary and link them just as the intermediary can. If a consumer does not consent to data storage using GDPR opt-out, then it immediately implies that certain types of cookies cannot be stored, thereby shutting the intermediary out. We can directly infer consumer privacy choices from the number of consumer visits as seen by this (third-party) intermediary and the change in composition, necessary to answer the first two questions. We also observe revenues from keyword-based online advertising, and observe the output of a proprietary machine learning algorithm that predicts the purchase likelihood, which will help us to address the third question.

Our empirical design exploits the fact that the intermediary contracts with many different platforms all around the world who were differentially impacted by the introduction of GDPR. Furthermore, the machine learning algorithm is trained and employed separately for each online travel website. This means that changes in data on one website, due to GDPR or other factors, do not impact the performance of the algorithm on other websites. We exploit these features of our data and the geographic reach of GDPR to utilize a difference-in-differences design for several outcome variables across major European countries and other countries where GDPR was not implemented.

We find that GDPR resulted in approximately a 12.5% reduction in total cookies, which provides evidence that consumers are making use of the increased opt-out capabilities mandated by GDPR. However, we find that the remaining set of consumers who do not opt out are more persistently trackable. We define trackability as the fraction of consumers whose identifier a website repeatedly observes in its data over some time period. We find that trackability has increased by 8% under GDPR.

We explore the mechanisms behind the increased trackability and argue that the most plausible explanation is that the individuals who make use of GDPR opt-out are primarily substituting away from other browser-based privacy means, such as cookie blockers, cookie deletion, and private browsing. However, GDPR opt-out and these other privacy means lead to a very different data generating process. Browser-provided privacy means assign a new ID to a consumer, thus making her appear as a new user, every time she visits the site. This results in many artificially short-lived consumers whereas GDPR privacy means simply remove these individuals from the data. As a result, those consumers that remain in the data after the implementation of GDPR are more persistently identifiable. We illustrate this difference in Figure 2.

Given this change in consumer composition, we explore the extent to which this affects advertising revenues. In our setting the revenues that we observe come from keyword-based advertising and, further, when consumers opt-out they are no longer exposed to advertisements. We

find that there is an immediate drop in the total number of advertisements clicked and a corresponding immediate decline in revenue. Over time, though, advertisers on average increase their bids for the remaining consumers, leading to a smaller overall decline in revenue. This indicates that the remaining set of consumers are higher value consumers compared to the pre-GDPR set of consumers. One possible mechanism for this is that the increased identifiability of consumers allows for advertisers to better attribute purchases to advertisements than before. This increased attribution ability leads to an increase in perceived overall value of consumers by advertisers.

Finally, we study the effect that GDPR had on the intermediary's ability to predict consumer behavior. In particular, we study the performance of the classifier used by the intermediary, which is a crucial element of its business. The classifier provides a prediction of the probability that a consumer will purchase on the website where she is currently searching. We find that there is evidence that the classifier did not immediately adjust to the post-GDPR distribution. However, despite this, we still find that the ability of the classifier to separate between purchasers and non-purchasers did not significantly worsen after GDPR and that, if anything, the changes to the data observed by the intermediary should lead to improvement in its ability to separate between purchasers and non-purchasers.

Related Work

The protection of consumer privacy and its consequences has been studied by economists, legal scholars, and computer scientists for several decades. We contribute to three strands of literature in the economics of privacy.

Consequences of Data Privacy Regulation: To the best of our knowledge Goldberg, Johnson and Shriver (2019) is the first paper to examine the economic impact of GDPR on European websites. They similarly adopt a difference-in-differences approach to study the effect of GDPR on the traffic and purchase volume of European websites. However, the scope of our data allows us to tie together the consequences of the opt out decisions of individuals with consumer identifiability, advertising revenues, and effectiveness of prediction technologies. This allows us to go beyond direct measurement of the impact of GDPR and further understand the externalities associated with individual privacy decisions and their indirect economic impact.

Several other papers have studied the impact of the GDPR in other domains. Jia, Jin and Wagman (2018) show that GDPR had adverse impacts on venture capital investment. Zhuo et al. (2019) studies the impact of GDPR on Internet connection agreements. Johnson and Shriver (2019) studies changes in market concentration of web technology vendors as a result of GDPR. Degeling et al. (2018) show that a significant number of websites responded to GDPR by updating their privacy policy statements and adding cookie consent, as was mandated by the policy. Utz et al.

(2019) showed that there was heterogeneity in the implementation of GDPR across websites that led to meaningful differences in whether consumers consented to data collection.

Further, several other papers study the effectiveness of previous data privacy regulations on online advertising. The first is the EU's 2009 ePrivacy Directive, also known as the Cookie Law, a previous European regulation aimed at increasing the transparency and control of consumer data. Goldfarb and Tucker (2011) use a survey-based methodology to study the effectiveness of online advertising in the EU after this law and find that advertising effectiveness dropped. The second is the self-regulated Ad-Choices program that allowed consumers to opt out of behavioral online advertising. Individuals who opt out under this program still see advertisements, but they are no longer targeted based on their personal histories. Johnson, Shriver and Du (2020) examine this policy and find that consumers that opt out generate 52% less revenue compared to consumers that do not, but that only less than a quarter of a percent of advertising impressions are from opt out consumers. Goldfarb and Tucker (2012*a*) argue that privacy regulations may hamper data-based innovation across a number of industries, including targeted advertising. Finally, Johnson (2013) estimates a structural model of advertising auctions and explores the effect of different hypothetical opt-in and opt-out data consent policies through counterfactual calculations. He finds that advertisement revenue would drop by 34.6% under an opt-in policy and by 3.9% under an opt-out policy.

Information Externalities: An important consequence of a consumer's privacy decision is the informational externality generated by that decision, as information revealed by one consumer can be used to predict the behavior of another consumer.[6,7] Several recent theoretical studies argue how such externalities can lead to the underpricing of data, and results in socially excessive data collection (Fairfield and Engel, 2015; Choi, Jeon and Kim, 2019; Acemoglu et al., 2019; Bergemann, Bonatti and Gan, 2019; Liang and Madsen, 2019). Braghieri (2019) theoretically studies how privacy choices by consumers can have pecuniary externalities on other consumers by affecting firms' incentives for price discrimination. The current paper identifies a novel form of informa-

[6]Implicit in the study of the effect on consumer predictability is the notion that privacy is not simply about the revelation of a consumer's information but also the ability of a firm to predict the behavior of a consumer. The idea that privacy is additionally a statistical notion is a common thread in the literature on differential privacy (Dwork, 2011; Dwork, Roth et al., 2014). Differential privacy studies the marginal value of an individual's data for the accuracy of a statistical query and gives a mathematical framework for trading off the privacy loss of an individual revealing her information and the marginal change in the accuracy of a statistical query. For a discussion of the economic mechanisms at play in differential privacy based methods, see Abowd and Schmutte (2019). While the intuition behind differential privacy is similar to what we study, we do not explore the design of algorithmic privacy tools. Rather, we empirically document the statistical consequences of privacy choices made by individuals on the predictability of others.

[7]There is also an emerging, broadly related, literature that studies implications of a more data-driven economy (Chiou and Tucker, 2017; Kehoe, Larsen and Pastorino, 2018; Decarolis and Rovigatti, 2019; Aridor et al., 2019; Bajari et al., 2019)

tional externalities. While the existing research focuses on how a consumer's decision to *reveal* her private data can predict the behavior of, and thus can inflict externalities on, those who *do not reveal* their data, we recognize externalites that run in the opposite direction. Namely, we show that the decision by a privacy-concerned consumer to switch from obfuscation to a more effective GDPR-enabled opt-out may increase the trackability of, and thus exert extenalities on, the opt-in consumers who *choose to reveal* their data. More importantly, to the best of our knowledge, this is the first paper that identifies privacy externalities empirically.

Preferences for Privacy: The broader literature on the economics of privacy, recently surveyed in Acquisti, Taylor and Wagman (2016), has studied the privacy preferences of individuals. One prevalent research strand is understanding the privacy paradox, which is the apparent disparity between stated and revealed preference for privacy. In particular, consumers state a strong preference for privacy, but are willing to give up their personal information for small incentives (Berendt, Günther and Spiekermann, 2005; Norberg, Horne and Horne, 2007; Athey, Catalini and Tucker, 2017). Acquisti, John and Loewenstein (2013) use a field experiment to evaluate individual preferences for privacy and find evidence of context-dependence in how individuals value privacy. Using stated preferences via a survey Goldfarb and Tucker (2012*b*) show that consumer's privacy concerns have been increasing over time. Lin (2019) shows via a lab experiment that consumer privacy preferences can be broken down into instrumental and non-instrumental components. Our study contributes to this literature by analyzing consumer privacy choices made in a consequential setting, instead of only looking at stated preferences. We find that a significant fraction of consumers utilize the privacy means provided by GDPR, giving suggestive evidence that consumers do value their privacy in consequential settings and not only say that they do.

The paper is structured as follows. Section 2 overviews the relevant details from European privacy law and consumer tracking technology. Section 3 describes the data and empirical strategy that is used for this study. Section 4 provides evidence on the degree to which consumers make use of the privacy tools provided by GDPR. Sections 5 and 6 analyze the extent to which this affects online advertising revenues and prediction, respectively. Section 7 concludes.

2 Institutional Details

In this section we discuss European privacy laws and the relevant details of the General Data Protection Regulation. We will then describe how websites track consumers online and how GDPR can affect such tracking.

2.1 European Data Privacy Regulation

GDPR was adopted by the European Parliament in April 2016. Companies were expected to comply with the new regulations by May 25th, 2018.[8] It required substantial changes in how firms store and process consumer data. Firms are required to be more explicit about their data retention policy, obligating them to justify the length of time that they retain information on consumers and delete any data that is no longer used for its original purposes. Furthermore, it required firms to increase the transparency around consumer data collection and to provide consumers with additional means to control the storage of personal data.

The primary component of GDPR that we focus on is the new data processing consent requirement. Under the regulation firms need *informed, specific, and unambiguous* consent from consumers in order to process their personal data, which requires consumers to explicitly opt into data collection. Recital 32 of the regulation spells out what consent means:

> *Consent should be given by a clear affirmative act establishing a freely given, specific, informed and unambiguous indication of the data subject's agreement to the processing of personal data relating to him or her, such as by a written statement, including by electronic means, or an oral statement. This could include ticking a box when visiting an internet website, choosing technical settings for information society services or another statement or conduct which clearly indicates in this context the data subject's acceptance of the proposed processing of his or her personal data. Silence, pre-ticked boxes or inactivity should not therefore constitute consent.*

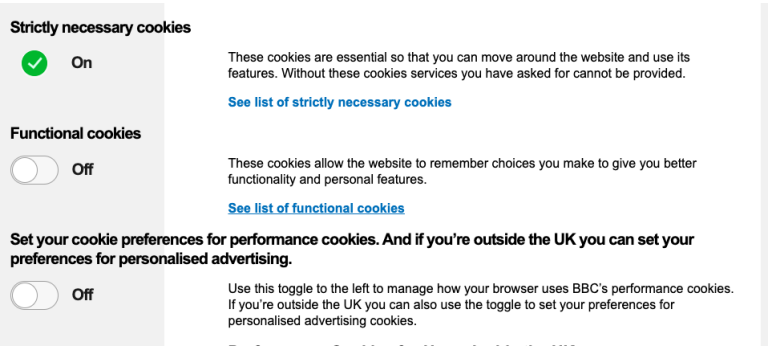
Panel (a) of Figure 1 shows an example of a post-GDPR cookie policy from the BBC, a news organization based in the United Kingdom, and panel (b) of Figure 1 shows a cookie policy of a firm in the United States. The former highlights the specifications of the law, specifying what type of cookies are stored for what purposes and giving consumers the opportunity to opt out from them individually. The latter has no explicit option for the consumers to opt out of data collection. Instead, it directs consumers to use browser-based privacy means, which allow to control the website's cookies.

The consent requirement is an important component of the law, though there were many other stipulations of the law that enhanced consumer privacy protection and required substantial changes by firms in order to be in compliance. The fines for non-compliance with the legislation are large - the maximum of €20 million, or 4% of total global annual sales for the preceding

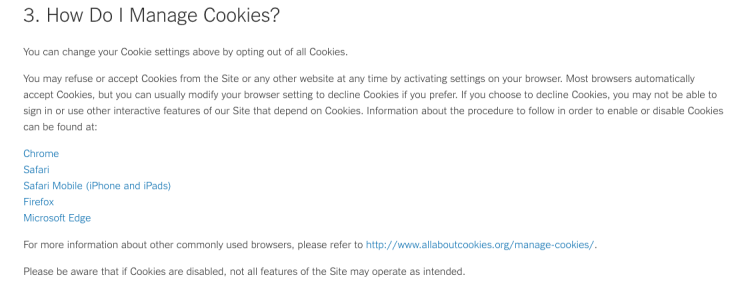
[8] GDPR was intended to overhaul and replace the Data Protection Directive which was enacted in 1995. GDPR further complements the other major European Privacy Regulation, The Privacy and Electronic Communications Directive, also known as the "Cookie Law". Relative to this law, GDPR strengthened the territorial scope to include data generated by EU consumers, no matter the location of the firm processing the data, and strengthened the degree of firm transparency and stipulations on consumer consent.

Figure 1: Example Consent Notifications

(a) Post-GDPR consent dialog



(b) Standard opt-out on US websites



Notes: The top panel shows a GDPR opt in consent dialog for the BBC. The dialog is explicit about the data that the website collects and requires the consumer to opt into all non-essential data collection. Each separate purpose of data processing is consented to individually. The bottom panel shows an "opt out" dialog for a website in the US that is not required to be GDPR compliant. The website directs consumers to manage their browser cookies and does not have any direct options for the consumer to opt out of data collection.

financial year - giving strong incentives for firms to comply with the regulation. According to PricewaterhouseCoopers, many firms are spending millions of dollars in order to comply with the regulation.[9] However, despite this observation, there was still considerable non-compliance around the onset of the law and in the next section we will discuss how this non-compliance affects the interpretation of our estimates.

[9]Pulse Survey: GDPR budgets top $10 million for 40% of surveyed companies. `https://www.pwc.com/us/en/services/consulting/library/general-data-protection-regulation-gdpr-budgets.html`. Retrieved on December 15th, 2019.

2.2 Consumer-Tracking Technology

The primary consumer tracking method that we focus on in this study are web cookies.[10] Cookies are small text files that are placed on consumer's computers or mobile phones. The attachment of a cookie gives websites, in principle, a persistent identifier. As long as the same cookie persists, they can attribute different sessions to the same consumer and, as a result, track them across time and different websites. However, privacy-conscious consumers can make use of various privacy means to control the degree of persistence of this identifier. The primary means available to them are browser-based tools, such as manually deletion of cookies, "private browsing" mode,[11] or cookie blockers.[12] These browser-based privacy means regenerate the cookie identifier but the data that is generated on the website is still sent and stored. The data is attributed to different consumers, even though they originate from the same consumer.

The GDPR opt-in rule provides another way for consumers to protect their privacy. The stipulations of GDPR, properly implemented and utilized by consumers, arguably provide a stronger protection than the aforementioned means since they block all non-essential information from being sent to the third-party website.[13] In our context, consumers should always have the option to opt out of having their data sent to the intermediary since it provides a non-essential, third-party service.

An important distinction for our purposes is, therefore, that browser-based privacy means do not prevent a consumer's data from being sent to the website and only give consumers control over the website's identifier. The consent requirement of GDPR goes one step beyond that and enables consumers to deny *any* data to be sent to the website. As a result, substitution between these two can lead to a different data generating process. Browser-based privacy means lead to many artificially short consumer histories, whereas GDPR opt-out simply removes the data completely.[14]

[10] Common alternatives are other forms of storage in the browser as well as device fingerprinting, which use Internet Protocol (IP) addresses combined with device specific information to identify individuals. However, these are less commonly utilized and importantly not utilized by the intermediary.

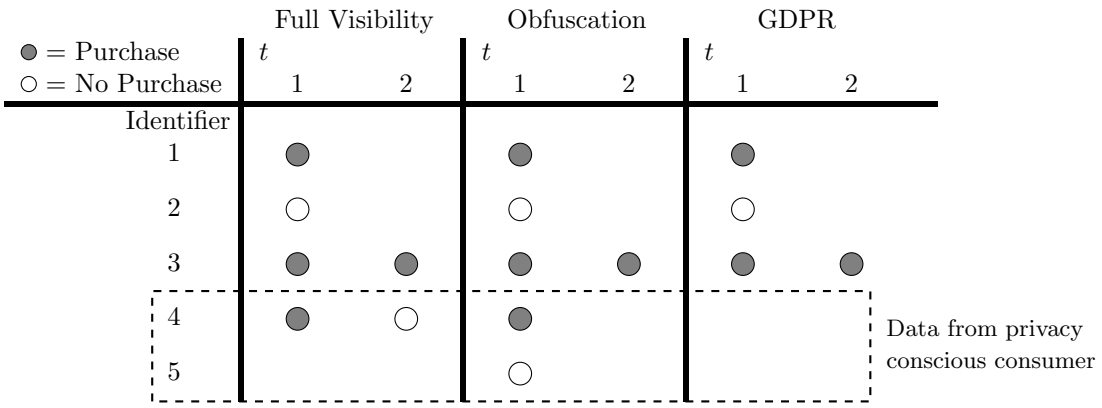
[11] Private browsing modes create "sandbox" browser environments where cookies are only set and used for the duration of the private browsing session. As a result, the website cannot link together data from the same consumer both before and after the private browsing section.

[12] There also exist industry opt-out services, such as the Ad Choices program, but these are relatively hard to use and have little usage (Johnson, Shriver and Du, 2020). Survey-based evidence informs us that the most utilized privacy means by consumers is manual cookie deletion (Boerman, Kruikemeier and Zuiderveen Borgesius, 2018)

[13] It is important to note that GDPR does not prevent "essential" information from being sent to a website. For instance, the ability to store consumer session cookies that allow them to provide a consistent consumer experience for the consumer may be considered "essential" information. The intermediary that we partner with, however, is a third-party service that provides complementary services to the primary functioning of the websites and so is not an "essential" service on any website where we observe data. As a result, any usage of GDPR opt-out shuts out data from being sent to the intermediary.

[14] It's important to point out that consumers can still make use of both privacy means and do not necessarily need to substitute from exclusively using browser-based privacy means towards exclusively using GDPR-provided

Figure 2: Illustration of Effects of Different Privacy Means on Data Observed



Notes: The leftmost column displays the identifier observed by the intermediary. The left panel represents the scenario where the behavior of each consumer is fully observable. The middle panel shows how, before GDPR, the privacy conscious consumer 4 has her identifier partitioned into two separate identifiers from the perspective of the intermediary. The right panel shows how, under GDPR, the data of the privacy conscious consumer, is not directly sent to the intermediary.

This is illustrated in Figure 2. The figure shows the data that is generated by four different consumers. "Full Visibility Baseline" shows a hypothetical scenario where each of the four consumers is fully identifiable. They generate spells of browsing sessions where each dot corresponds to one session and the color of the dot indicates whether or not the consumer purchased a good on the website as a result of that search. Suppose that only consumer four is privacy-conscious. Before GDPR, consumer four can protect her privacy by deleting her cookies and regenerating her identifier. This is illustrated in the second panel of the figure where the two sessions for this consumer are associated with two separate identifiers from the perspective of the intermediary. However, the third panel shows that, when GDPR opt-out is available, this consumer opts out and their data completely disappears.

The figure also illustrates how the different data scenarios impact consumer predictability and how the privacy choices of individuals lead to informational externalities. The intermediary's objective is to predict the probability that an identifier will purchase the next time that she appears on the website. The intermediary has available the full history associated with the identifier so that she can design a prediction rule that depends on how many times this identifier has appeared on the website and how often this has resulted in purchase. Under full visibility, each consumer has a distinct search and purchase history so that the intermediary gets a distinct

privacy means. However, from the perspective of the intermediary and websites in general, once a consumer utilizes GDPR opt-out then, since they no longer see any data from this consumer, the browser-based privacy means become irrelevant. As a result, from their perspective, it appears as a direct substitution.

signal and can adjust its prediction rule accordingly. However, under obfuscation, consumer four deletes her cookies and gets partitioned into two separate identifiers. Now the consumer history associated with identifier 4 is identical to that of identifier 1 and the consumer history associated with identifier 5 is identical to that of identifier 2. Thus, the ability of the intermediary to predict the behavior of consumers 1 and 2 is affected by the privacy protection employed by consumer 4 as, even though their histories are not the same, they appear so to the intermediary. Under GDPR, on the other hand, consumer 4's data is not observed at all. While this leads to a loss in the amount of data, it removes the externality that consumer 4 imposed on consumer 1 and 2 that occurred under obfuscation and improves the ability of the intermediary to predict their behavior.

3 Data and Empirical Strategy

We obtained access to a new and comprehensive dataset from an anonymous intermediary that records the entirety of consumer search queries and purchases across most major online travel agencies (OTAs) in the United States and Europe as well as most prominent travel meta-search engines from January 1st, 2018 until July 31st, 2018. We observe consumer searches, online advertising, the intermediary's prediction of consumer behavior, and consumer purchases.

3.1 Data Description

The disaggregated data contains each search query and purchase made on these platforms as well as the associated advertising auction for each query. In a single search query the data contains: the identifier of the consumer, the time of the query, the details of the query (i.e. travel information), an identifier for the platform, the browser, the operating system, and the estimated probability of purchase on the website according to the predictive machine learning algorithm employed by the intermediary. For a subset of the websites, we observe purchase information containing the consumer identifier and time of purchase.

Each query can trigger an advertising auction. In that case, the data contains: the number of bidders in the auction, the values of the winning bids, and an identifier for the winning bidders. Furthermore, if a consumer clicks on the resulting advertisement, the click itself and the resulting transfer between the advertiser and the intermediary are recorded.

Our analysis utilizes an aggregation of this dataset by week, operating system, web browser, website identifier, and country. The data was aggregated on a weekly level to remove unimportant day-of-the-week fluctuations. Further, the GDPR compliance date was May 25th, 2018, which was on a Friday and, as a result, our data was aggregated on a Friday-to-Friday level. Note that the

GDPR compliance date corresponds to the beginning of the 22nd week in the year according to our labeling.[15]

3.2 Empirical Strategy

To understand the causal effect of GDPR we rely on a difference-in-differences design that exploits the geographic reach of the EU GDPR regulation. The regulation stipulates that websites that transact with EU consumers were required to ask consumers for explicit consent to use their data through an opt-in procedure, while those who processed non-EU consumers data were not obligated to do so. Even though many online travel companies transact with consumers in several countries around the world this specification works well in our setting since it is common for online travel websites to have separate, country-specific, versions of their websites and only the websites intended for EU countries are made GDPR compliant.

Our analysis focuses on the effect of the overall policy and not the effect of specific implementations of the policy. Thus, the treatment date of the policy corresponds to the GDPR compliance date, which was May 25th, 2018 (or the beginning of week 22). Our treatment group consists of nearly the universe of travel websites in major EU countries (at the time): Italy, the United Kingdom, France, Germany, and Spain. Our control group consists of nearly the universe of travel platforms in the United States, Canada, and Russia. These countries were chosen as controls since EU laws do not directly apply to them, but their seasonal travel patterns are similar to those in the EU countries as a result of similar weather and vacation patterns throughout the year.

Our primary regression specification is the following for the outcome variables of interest where c denotes country, j denotes the website, o denotes operating system, b denotes web browser, p denotes product type (hotels or flights), and t denotes the week in the year:

$$y_{tcjobp} = \alpha_t + \delta_{jc} + \kappa_c + \xi_j + \gamma_o + \zeta_b + \omega_p + \beta(EU_j \times after) + \epsilon_{tcjobp} \tag{1}$$

EU_j denotes a website subject to the regulation, $after$ denotes whether the current week is after the GDPR compliance date (i.e. week 22 or later), α_t denotes time fixed effects, δ_{jc} denotes country-specific website fixed effects, κ_c denote country fixed-effects, ξ_j denotes website fixed effects, ω_p denotes product type fixed effects, γ_o denotes operating system fixed effects, and ζ_b denotes browser fixed effects. Our standard errors are clustered at the website-country level.

In order to validate parallel trends and to understand the persistence of the treatment effect,

[15]Note that we further enforce a balanced panel by dropping any observation that has zero logged searches in any period during our sample period. We do this in order to ensure that our results are not biased from entry / exit of websites into our data during the sample period, which would bias our estimates. To our knowledge, this entry and exit is usually a result of varying contractual relations between the intermediary and the websites and so is largely orthogonal to our variables of interest.

we further utilize a regression specification that captures the potentially time-varying nature of the treatment:

$$y_{tcjobp} = \alpha_t + \delta_{jc} + \kappa_c + \xi_j + \gamma_o + \zeta_b + \omega_p + \sum_{k=\underline{T}}^{\bar{T}} \beta_k EU_j + \epsilon_{tcjobp} \tag{2}$$

The variable definitions are the same as before and we similarly cluster our standard errors at the website-country level.

We run our regressions over the time period between weeks 16 and 29 of 2018, which is between April 13th and July 20th. The GDPR compliance date aligns with the beginning of week 22. Further, week 20 is consistently the baseline week in our regressions since there are some firms that began to implement GDPR near the end of week 21 and so week 20 is the last week where there should be no direct impact from GDPR as a result of website implementation.[16]

Our empirical strategy centers around the official GDPR implementation date. However, each website had to individually implement the changes stipulated by GDPR and there is evidence that there was considerable heterogeneity in compliance among firms. Further, even within the subset of firms that complied with the regulation, the degree to which consumers responded varied considerably based on the nature of implementation (Utz et al., 2019). As a result, we would want to include information on the timing and degree of implementation across the various websites in our sample. However, due to technical limitations, we cannot directly observe the timing and degree of GDPR implementation during the time period we study.[17]

Thus, any effects that we observe with our empirical specification are a combination of the explicit consequences as a result of implementing the stipulations of GDPR for the subset of websites that implemented it and any changes in advertiser and consumer behavior in response to the increased saliency of privacy considerations on the Internet.[18] Thus, since we do not observe the full extent of non-compliance, our estimates can be viewed as a lower bound on the true impact of the policy had websites all fully complied with it.

[16] Our dataset ends on July 31st, 2018, which is a Tuesday, and an important measure that we want to track is the amount of consumer persistence on a weekly level, which looks at the fraction of observed cookies that remain observable in the data after some number of weeks. Since this measure requires a complete week of data to compute properly, we drop the incomplete week at the end of July as well as the full last week in July so that we can have consistency between the regressions on aggregate consumer response and those on consumer persistence.

[17] We attempted to utilize tools such as the Wayback Machine, which takes snapshots of websites across the entire Internet frequently. However, the coverage of relevant websites on the Wayback Machine is spotty and, given that many of the consent dialogs for GDPR consent are dynamically generated, are not always picked up by the snapshot taken of the website.

[18] It would be interesting to isolate the effects of each possible channel, though our data limitations prohibit us from doing so. We were able to verify that several websites in our sample implemented GDPR consent guidelines around the time of the policy and that several websites in our sample did not, though there are a considerable number for which we are uncertain when they implemented the policy.

4 Consumer Response to GDPR

In this section we quantify the extent to which consumers utilize the GDPR-mandated ability to opt out. We measure how GDPR opt-out impacts the total number of cookies and searches observed by the intermediary. We then explore whether there were any changes in the composition of the remaining, opted-in consumers.

4.1 Opt-Out Usage

Recall that we do not directly observe opt-out in our dataset because consumers who opt out are no longer part of our dataset. As a result, at time t, the total number of consumers on a website j is given by the true number of consumers subtracted by the number of consumers who have opted out.

$$U_{jt}^{OBS} = U_{jt}^{TRUE} - U_{jt}^{OPT-OUT}$$

In the control group, $U_{jt}^{OPT-OUT} = 0$, whereas post-GDPR $U_{jt}^{OPT-OUT} \geq 0$. We assume parallel trends in U_{jt}^{TRUE}, which means that any change in U_{jt}^{OBS} allows us to identify $U_{jt}^{OPT-OUT}$.[19,20]

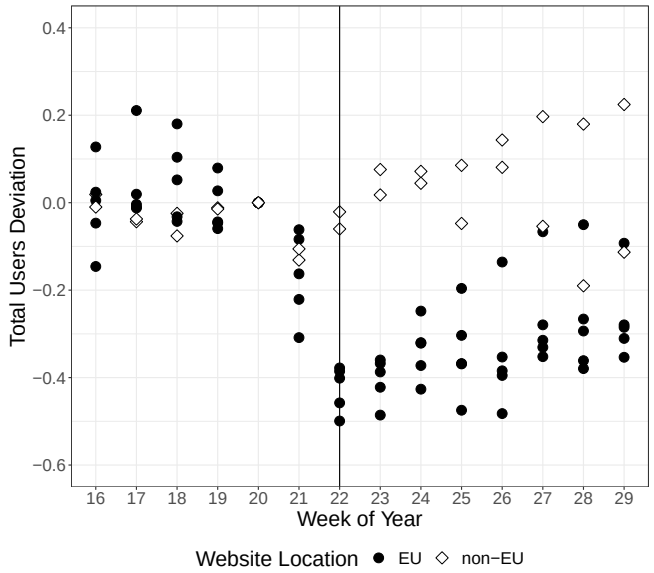
Figure 3 displays the total unique cookies over our sample period for a single multi-national website close to the implementation date and shows a clear drop at the onset of GDPR. Columns (1) and (2) of Table 1 report the result of regression (1) with total number of observed unique cookies as the outcome variable. We consider the specification in both levels and logs. The estimates show that, in aggregate, GDPR reduced the total number of unique cookies by around 12.5%.

It is important to note that this result *does not* imply that 12.5% of consumers made use of the opt-out features. This is because the unit of observation is a cookie, rather than a consumer. A single consumer can appear under multiple cookie identifiers if they make use of the aforementioned browser-based privacy means. Nonetheless, the results point to a relatively large usage of the opt-out features by consumers.

[19] As noted in Goldberg, Johnson and Shriver (2019), another possible complication is that this could be a result of firms changing the type of data that they send to third party services. To our knowledge there is no change in the data the websites send to the intermediary as a result of GDPR since the intermediary and the data are crucial for generating advertising revenue for these websites. Further, if a website decided to stop using the intermediary altogether then, as noted previously, they would not be part of our sample.

[20] Another possible confounding factor is the sales activity of the intermediary. For instance, it's possible for the intermediary to sell additional advertising units to a website that can appear on pages of the website where the intermediary previously was not tracking before. If there was a differential effect from this around the date of the treatment then this could systematically bias the number of unique cookies and searches that we observe. To test the plausibility of this hypothesis, we run our difference-in-differences specification with these outcome variables with the results in Table 5 showing that there was no significant change. Thus, we rule out this as a possible explanation.

Figure 3: Total Number of Unique Cookies for a Single Multi-National Website.



Notes: A single point on the graph represents the total number of unique cookies for a single country. The reported value on the y-axis is percent deviation relative to week 20, or $\frac{U_t - U_{t=20}}{U_{t=20}} \quad \forall t \neq 20$

Table 1: Difference-in-Differences Estimates for Cookies and Searches

	(1) log(Unique Cookies)	(2) Unique Cookies	(3) log(Recorded Searches)	(4) Recorded Searches
DiD Coefficient	-0.125**	-1378.1*	-0.107*	-9618.3**
	(-2.43)	(-1.71)	(-1.87)	(-2.24)
Product Type Controls	✓	✓	✓	✓
OS + Browser Controls	✓	✓	✓	✓
Week FE	✓	✓	✓	✓
Website × Country FE	✓	✓	✓	✓
Observations	63840	63840	63840	63840

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: t-statistics are reported in parentheses. The standard errors for every regression are clustered at the website-country level. We aggregate every dependent variable to the website-country-product type-week level between weeks 16 and 29, including both weeks 16 and 29 (April 13th - July 20th). The dependent variables in the regression reported in the first and second column are the log and overall level of the number of unique cookies observed. The dependent variables in the regression reported in the third and fourth column are the log and overall level of the number of total recorded searches.

Another measure of consumer response is the total number of searches that is recorded by the intermediary. We re-run the same specification with recorded searches as the dependent variable and report the results in columns (3) and (4) of Table 1. We find that there's a 10.7% drop in the overall recorded searches which is qualitatively consistent with the effect size of the specification using the number of unique cookies.

In order to demonstrate the validity of the difference-in-differences estimates for both of these outcome variables, we provide evidence that the parallel trends assumption holds in our setting by estimating the time-varying treatment specification (2). Figure 8 shows the resulting treatment effect over time and points to parallel trends being satisfied as well as a consistent treatment effect size over our sample period.

4.2 Persistence of Identifier

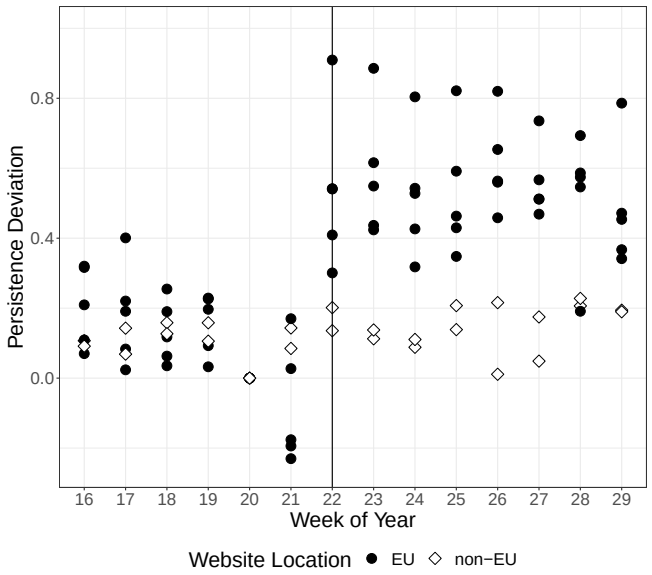
A natural question is whether there were any changes in the composition of the remaining consumers who did not opt out. Our primary measure for investigating this is by tracking the average persistence of a consumer identifier before and after GDPR. We define an *identifier persistence* measure that tracks how often cookies that we see in a given week return after k weeks, where we explore different values for k (1,2,3, and 4 weeks). Let C_{jt} be the set of cookies seen in week t on website j, the measure is then given by:

$$persistence_{kt} = \frac{|C_{j,t} \cap C_{j,t+k}|}{|C_{j,t}|}$$

In Figure 4 we set $k = 1$ and display the persistence measure for a single multi-national website with country-specific versions of the website over time. At the onset of GDPR there is a clear increase in persistence on the EU-based websites, but no noticeable difference in the non-EU websites. We further validate this increase by running our baseline difference-in-differences specification using the persistence outcome variable for $k \in \{1, 2, 3, 4\}$.[21]

[21]Note that in order to run specification (1) we drop the last 4 weeks of our sample so that we are utilizing the same sample as we vary k. However, our results are qualitatively robust to including these weeks when the data for them is available.

Figure 4: One Week Persistence for a Single Multi-National Website



Notes: A single point on the graph represents the one week persistence fraction for a single country. The reported value on the y-axis is percent deviation relative to week 20, or $\frac{persistence_{1,t} - persistence_{1,t=20}}{persistence_{1,t=20}} \quad \forall t \neq 20$

Table 2: Difference-in-Differences Estimates for Consumer Persistence

	(1) 1 Week Persistence	(2) 2 Weeks Persistence	(3) 3 Weeks Persistence	(4) 4 Weeks Persistence
DiD Coefficient	0.00308*	0.00416***	0.00382***	0.00505***
	(1.96)	(3.40)	(3.10)	(3.50)
Product Type Controls	✓	✓	✓	✓
OS + Browser Controls	✓	✓	✓	✓
Week FE	✓	✓	✓	✓
Website × Country FE	✓	✓	✓	✓
Observations	50160	50160	50160	50160

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: t-statistics are reported in parentheses. The standard errors for every regression are clustered at the website-country level. We aggregate every dependent variable to the website-country-product type-week level between weeks 16 and 26, including both weeks 16 and 26 (April 13th - June 29th). The dependent variables in the regression are the consumer persistence measures for $k = 1, 2, 3, 4$, respectively.

Table 2 shows the results of this regression, which indicate that there is a statistically significant and meaningful increase in consumer persistence and that this effect gets more pronounced as k increases.[22] We further run the time-varying treatment specification (2) in order to validate that parallel trends holds and to understand the consistency of the effect over time. Figure 9 shows that while for $k = 1$ the time dependent treatment effects are more noisy, for all $k \geq 2$ parallel trends hold and the treatment effect is stable over time.[23] The treatment effect remains roughly the same as k grows, even though Table 6 shows that the mean persistence declines as k increases. For instance, in the pre-treatment period, the mean persistence for EU websites was 0.0597 and the estimated treatment effect is 0.005 indicating a roughly 8% increase in persistence as a result of GDPR.

The economic implications of the increase in persistence depend on the mechanisms that drive this effect for which we have two plausible hypothesis. The first is a *selective consent hypothesis* where consumers only consent to data processing by websites that they frequently use. Under this hypothesis, if infrequent users of a website are those who deny consent to data storage more-so than frequent users, then the remaining set of consumers will naturally appear to be more persistent. The second is a *privacy means substitution hypothesis* where privacy conscious consumers who were previously making use of brower-based privacy means now utilize GDPR opt-in to protect their privacy. Recall that the utilization of these privacy means would result in many artificially short-lived consumers. If these same consumers utilize GDPR opt-in instead they would no longer show up in the intermediary's dataset and the remaining set of consumers would appear to be more persistent even though their true search and purchase behavior may not have changed.

If the selective consent hypothesis is the predominant explanation for the increased persistence, then privacy regulation may favor firms with more established reputations or offer a wider variety of services.[24] The hypothesis would imply that in the long run consent for data collection can serve as a barrier to entry for newer firms with less established reputations and a smaller variety of services.

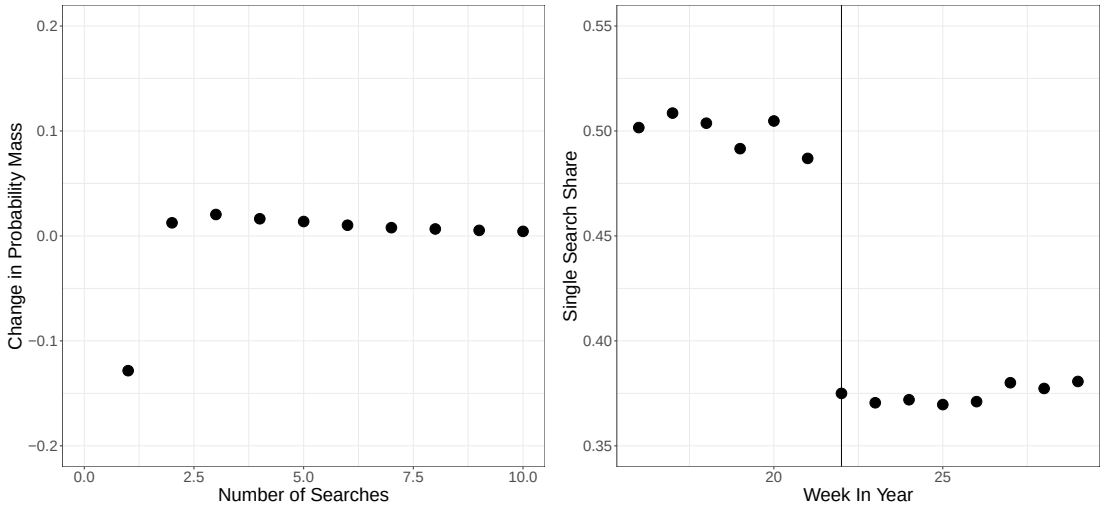
If the privacy means substitution hypothesis is the predominant explanation for the increased

[22] It is important to note that the persistence measure may have some noise when $k = 1$ due to consumer activity near the end of the week that spills over into the next week and falsely appears as persistence. As a result, the most reliable measures of consumer persistence are for $k \geq 2$, but we report $k = 1$ for completeness.

[23] Further, Figure 10 in the appendix shows the overall distributions of consumer persistence for the EU vs. non-EU and note that there are some outliers. In particular, there is a large mass of high persistence observations and persistence measures close to 0. Our results are qualitatively robust to running our specifications winsorizing and dropping these observations as well.

[24] There is a connection of this hypothesis to the theoretical predictions in Campbell, Goldfarb and Tucker (2015), who argue that consent-based data collection practices would allow larger firms to collect more data than smaller firms since they offer a wider scope of services. As a result, consumers may utilize these websites more and trust the website with their data more.

Figure 5: Change in Search Distribution for One Site



Notes: The figure on the left shows the difference in the share of consumers with x searches in the full sample after GDPR compared to before GDPR. For instance, the leftmost point indicates that there was a roughly 12.8% decrease in the share of cookies associated with a single search. The figure on the right breaks down the share of cookies associated with only one search week by week, as opposed to pooling the full sample periods before and after GDPR.

persistence, then there are several economically relevant consequences. First, the benefit of GDPR would be the marginal effect over existing privacy protection. Second, the usage of GDPR opt-out would lead to an externality on the opt-in consumers and, as a result, their privacy protection may be weakened. This would directly mean that firms relying on prediction may not suffer as much as the number of opt out indicates since this would enhance their prediction capabilities. Finally, it would allow for better advertisement attribution and measurement of advertising effectiveness which would directly influence the price advertisers are willing to pay for advertising.

We provide suggestive evidence that the privacy means substitution hypothesis is the more plausible of the two. We focus on one large hotels website in Germany and study the distribution of number of searches per cookie on this website. While both hypotheses imply that the drop in relative probability mass should be concentrated towards the lower end of the support, one predominant signature of browser-based privacy protection is a large mass of "single search" consumers. This comes from consumers who utilize cookie blockers, which results in continual regeneration of cookies after every request and leads to a large number of artificially short-lived consumers with only one logged search. Thus, we track the overall share of consumers with one search over a week in Figure 5. Consistent with the privacy means substitution hypothesis, we observe a discontinuous drop in the share of single-search consumers at the advent of GDPR suggesting that these consumers substituted towards utilizing GDPR opt-out.

We further study the overall distribution of consumer searches before and after GDPR. Fig-

ure 5 shows that only the probability mass of single searchers seems to have dropped and the shift in probability mass seems to be roughly evenly distributed across the different number of searches. This provides strong evidence that, at least for this site, the increase in persistence is largely driven by the drop in the "single searchers" which is consistent with what we would expect under the privacy means substitution hypothesis. Under the selective consent hypothesis we would expect that the loss in probability mass would be more evenly distributed across searches.

Finally, in order to provide additional evidence for the privacy means substitution hypothesis, we estimate heterogeneous treatment effects across web browsers and operating systems. While the selective consent hypothesis should imply no differences across these dimensions, the privacy means substitution hypothesis is more plausible for web browsers and operating systems with more technically sophisticated users and weaker existing privacy protections. Thus, we should expect a larger increase in persistence on these web browsers and operating systems, which is consistent with the results we find. The results and a full discussion are deferred to Appendix B.

5 GDPR and Online Advertising

We now study the effect of GDPR on the market for online advertising. Specifically, we investigate the extent to which the consumer opt-out and increase in average trackability affected advertiser's average value of consumers and overall revenue for both advertisers and websites. Advertisements are sold via real-time auctions.[25] Bidding does not occur at the level of individual consumer profiles but instead at the keyword segment level. An example of a keyword segement is the collection of searches for flights from JFK to LAX. Thus, any changes in advertiser behavior would result from changes in their overall estimate of the value of consumers. Bids are submitted per click and a payment from the advertiser to the intermediary only occurs if the consumer clicks on the advertisement.

First, we use a difference-in-differences specification to investigate whether the drop in unique identifiers and searches was associated with a similar drop in the total number of advertisements served.[26] Table 9 shows a drop in the number of advertisements, but the drop is not statistically significant. Figure 11 shows the time varying treatment effect, which confirms this pattern. This shows that the drop in advertisements shown is not as stark as the drop in unique cookies and recorded searches.

[25] The auction format is a linear combination of a generalized first and second price auction where there are N advertisers and k slots.

[26] This relationship is not necessarily mechanical since the individuals who opt out may differ on other dimensions, such as their propensity to click on advertisement and purchase.

Table 3: Difference-in-Differences Estimates for Advertising Outcome Variables

	(1) asinh(Total Clicks)	(2) Total Clicks	(3) asinh(Distinct Clicks)	(4) Distinct Clicks	(5) asinh(Revenue)	(6) Revenue	(7) Average Transfer	(8) Average Bid
DiD Coefficient	-0.135**	-251.9*	-0.133**	-214.9*	-0.168	-32972.3	28.97**	15.41***
	(-2.32)	(-1.91)	(-2.33)	(-1.84)	(-1.54)	(-0.75)	(2.12)	(2.90)
OS + Browser Controls	✓	✓	✓	✓	✓	✓	✓	✓
Product Category Controls	✓	✓	✓	✓	✓	✓	✓	✓
Website× Country FE	✓	✓	✓	✓	✓	✓	✓	✓
Week FE	✓	✓	✓	✓	✓	✓	✓	✓
Observations	62328	62328	62328	62328	62328	62328	62328	62328

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: t-statistics are reported in parentheses. The standard errors for every regression are clustered at the website-country level. We aggregate every dependent variable to the website-country-product type-week level between weeks 16 and 29, including both weeks 16 and 29 (April 13th - July 20th). The dependent variable in the regression reported in the first column is the total number of clicks associated with each observation and the second column is the inverse hyperbolic sine transform of this value. Likewise, the dependent variables in the third and fourth columns are the total number and inverse hyperbolic sine transform of the total number of unique cookies who interacted with advertisements. The dependent variables in the fifth and sixth column are the total number and inverse hyperbolic sine transform of the total revenue. The dependent variable in the seventh column is the average transfer between the intermediary and advertisers and in the eighth column it is the average bid by advertisers. Since some of the outcome variables can take zero values, in order to preserve these observations we utilize a common transformation in the applied microeconomics literature and use the inverse hyperbolic sine transform instead of the natural logarithm of the outcome variables (Bellemare and Wichman, 2019). The resulting transformed outcome variable, $\bar{y}$, is given by $\bar{y} = arcsinh(y) = \ln(y + \sqrt{y^2 + 1})$

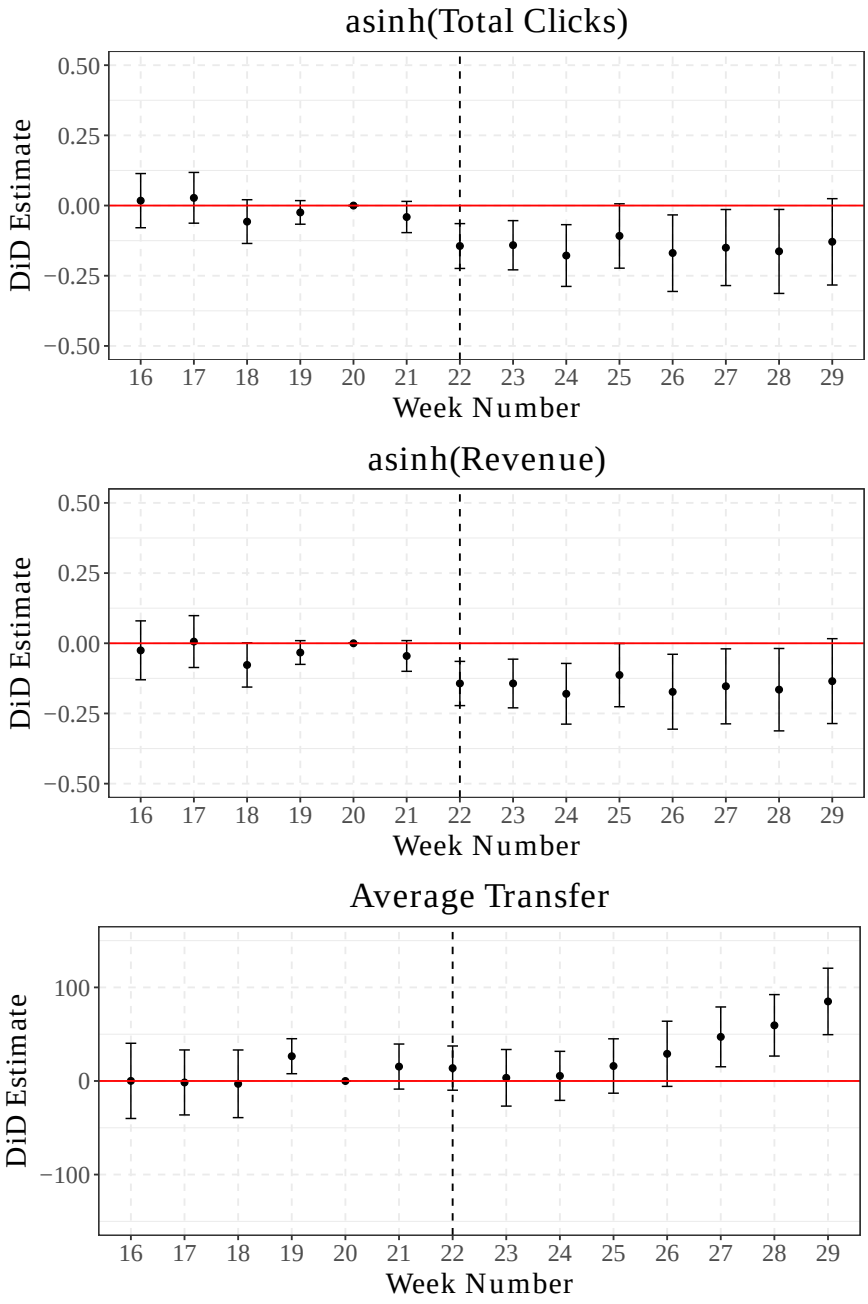
We next examine the effect on overall revenues earned by websites and advertisers. Revenues depend both on the number of clicks as well as the price per click. Columns (1) - (2) of Table 3 show that there is a statistically significant decrease of 13.5% in the total number of clicks, with an effect size commensurate with the drop in total cookies and searches. Furthermore, we look for changes in the number of distinct cookies associated with a click to see if any changes were driven by some small set of consumers who drive advertising revenues. Columns (3) - (4) show that the number of distinct clicks also decreases significantly. Finally, Figure 6 displays the time-varying specification for these outcome variable and shows that the effect on the number of clicks is relatively constant after the GDPR implementation date.

Columns (5) and (6) of Table 3 provide the estimate of the impact on revenue, which is negative though not statistically significant. The time-varying treatment effect displayed in Figure 6 shows that revenue initially falls sharply after the implementation of GDPR and then begins to increase. Importantly, columns (7) and (8) of Table 3 show that bids and average transfers from advertisers to the intermediary *increase*. We interpret this as an increase in the perceived average value that advertisers place on remaining consumers after GDPR.

Figure 6 displays the time-varying coefficient for the average transfer for a consumer and shows that the transfer does not change initially after the policy and then increases gradually. As a result, the immediate drop in clicks following GDPR leads to an immediate drop in revenue, but the increase in the average transfer for consumers after GDPR leads to a recovery of some of the lost revenue for the intermediary and advertisers.

We now explore the mechanisms behind the increase in bids for consumers. One possibility is that just as GDPR was implemented there was also a change in the composition of advertisers. Advertisers could have have entered or exited from the European market as a result of GDPR or for other unrelated reasons. In order to explore whether this was the case we compute the share of winning bids for each advertiser and track any changes in market concentration. We compute the two most commonly utilized market concentration measures: Concentration Ratio and the Herfindahl-Hirschman Index. We utilize our previously defined specifications to see whether there was any change in market concentration. The details of this exercise are described in Appendix D. We find a statistically significant but economically small increase in bidder concentration. Increased concentration in advertisers should have decreased the prices, not increased them. Therefore, we conclude that changes on the demand side are likely not the reason for the observed price increase.

Figure 6: Week by Week Treatment Effect for Total Clicks, Revenue, and Average Transfer



A more plausible explanation relates directly to the result that remaining consumers are now more trackable and therefore advertisers are better able to attribute purchases to advertisements. Advertisers assess the value of a consumer based on the conversion rate of an advertisement, which is the fraction of consumers that end up purchasing a good after clicking on an advertisement. However, effective measurement requires that the advertiser and intermediary can attribute purchases to advertisements, which relies on their capability to track consumers across time and websites. For instance, if a consumer deletes her cookies after clicking the advertisement and then subsequently purchases a good, the advertiser and intermediary will not be able to attribute the advertisement to a purchase. An implication of this is that consumers that are utilizing browser-based privacy means bias the perceived value advertisers have of consumers downwards. If the consumers that are utilizing these privacy means alternatively utilize the privacy means offered by GDPR instead, they will not appear in the sample of the advertisers whatsoever, which will lead the advertisers to gradually increase their perceived value of consumers.

As an illustrative example, suppose that there are five consumers who click on an advertisement. Suppose one of them (from hereon consumer A) deletes her cookies but ends up purchasing and, from the remaining four, suppose two of them end up purchasing. Thus, regardless of the behavior of consumer A, the advertiser's estimated conversion rate is 0.4, which is only correct in the case where consumer A never purchases. Suppose, instead, that GDPR opt-out is available and consumer A is removed from the sample of the advertiser and therefore never clicks on an advertisement. The advertiser's estimated conversion rate is 0.5 now, as opposed to 0.4 and so the perceived value of consumers weakly increases irregardless of consumer A's true behavior. More generally, dropping individuals similar consumer A from the observed sample can only weakly increase the advertiser's perceived value.

The changes that we observe in the advertising market are, therefore, consistent with the observation in Section 4. These results strongly suggest that GDPR has made it easier for advertisers to assess the value of consumers by improving their ability to track them and measure the effectiveness of advertisements.

6 GDPR and Prediction of Consumer Behavior

In this section we investigate whether the changes due to GDPR have affected the intermediary's ability to predict consumer behavior. Based on our analysis we expect there to be three predominant reasons why we might observe a change in the ability to predict. First, GDPR has significantly reduced the overall amount of data. Second, remaining consumers have longer histories and are more trackable. Third, in line with our illustration in Figure 2, GDPR might reveal correlation structures between between consumer behavior and the length of consumer histories

that were previously obfuscated by the use of alternative privacy tools. We would expect the first effect to decrease predictive performance and the second and third to increase predictive performance.

We take as given both the setup of the prediction problem and the algorithm that the intermediary uses. This allows us to understand the effects of GDPR on the prediction problem "in the field." Its problem is to predict *whether a consumer will purchase from a site she visits* based on history that the intermediary observes about this consumer. Specifically, its algorithm classifies a search by a consumer into two categories: *purchasers* and *non-purchasers*, based on whether the consumer will purchase a product on the current website *within some time window*. Formally, each query is classified into

$$y_{ijk} = \begin{cases} 1, & \text{if } i \text{ is a purchaser on website } j \text{ after search } k \\ 0, & \text{if } i \text{ is not a purchaser on website } j \text{ after search } k, \end{cases}$$

for a cookie i on website j on the kth query observed by the intermediary. For every cookie i we observe a series of searches on website j, $X_{ij1}, X_{ij2}, ..., X_{ijn}$ and, if the consumer ended up making a purchase on this website, the time-stamp of when consumer i purchased on website j. For every X_{ijk} that we observe, $y_{ijk} = 1$ if the purchase occurs within N days of the query and $y_{ijk} = 0$ otherwise. While in practice the value of N is platform-dependent, for our analysis we restrict focus to $N = 2$.[27] We will denote the *class proportion* as the proportion of searches that are associated with the purchaser class.

For each search, the intermediary produces a probability estimate that the consumer is a purchaser:

$$p_{ijk} = \Pr(y_{ijk} = 1 \mid X_{ij1}, ..., X_{ijk}), \forall i, j, k \tag{3}$$

We observe the intermediary's predicted $\hat{p}_{ijk}$ for every search.

In practice, for its own operation, the intermediary classifies each consumer into the two groups in real time during each search. This determination is based on whether the consumer's "score" $\hat{p}_{ijk}$ is above or below a chosen threshold $\hat{P}$.[28]

[27] For the majority of websites in our sample the intermediary sets $N = 1$ or $N = 2$. Further, from our preliminary analysis, the results do not qualitatively differ between $N = 1$ and $N = 2$.

[28] The value of this threshold is determined by revenue considerations and other factors irrelevant to the quality of predictions. As a result, we restrict focus on the probabilistic estimate $\hat{p}_{ijk}$ and not the *actual* classification.

6.1 Prediction Evaluation Measures

To evaluate the performance of the classifier deployed by the intermediary, we use two standard measures from the machine learning literature: the *Mean Squared Error (MSE)* and *Area under the ROC Curve (AUC)*.[29]

The MSE computes the mean of the squared errors associated with the predicted estimate $\hat{p}_{ijk}$ relative to the realized binary event. Specifically, let $\mathcal{I}_j$ be the set of all consumers on website j and let $\mathcal{K}_{ij}$ be the set of all events for consumer i on website j. Then, the MSE of website j is given by,

$$MSE_j = \frac{1}{\sum_{i \in \mathcal{I}_j} |\mathcal{K}_{ij}|} \sum_{i \in \mathcal{I}_j} \sum_{k \in \mathcal{K}_{ij}} (\hat{p}_{ijk} - y_{ijk})^2, \tag{4}$$

with a low MSE indicating a good prediction performance.

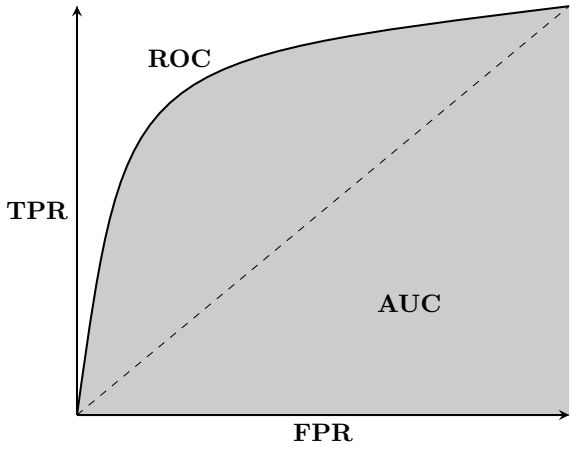
Although commonly used, the MSE has a couple of drawbacks for the current purpose. First, the measure is sensitive to the skewness of, and the change in, the class distribution. In the current context, about 90% of the searches result in non-purchase, which means that the estimate $\hat{p}_{ijk}$ tends to be low; intuitively, the estimate would tolerate more errors associated with the "infrequent" event (purchase) in order to minimize the errors associated with the more "frequent" event (non-purchase). Suppose now the class distribution changes so that more searches result in purchases. This is indeed what happens in our data after GDPR. Then, even though the consumer may not have become less predictable, MSE would rise artificially, due to the convexity associated with the formula, especially if the prediction algorithm does not adjust to the change in the distribution. Second, perhaps not unrelated to the first issue, the MSE is not the measure that the intermediary focuses on for its operation as well as for communicating with its partners. Instead, it focuses on AUC (the area under the curve), which we now turn to.

The AUC measures the area under the Receiver Operating Characteristic (ROC) curve.[30] The ROC curve in turn measures how well the classifier trades off Type I ("false positive") with Type II ("false negative") errors. To begin, fix the classification threshold at any $\hat{P}$. Then, a consumer with score $\hat{p}_{ijk}$ is classified as a purchaser if $\hat{p}_{ijk} > \hat{P}$ and a non-purchaser if $\hat{p}_{ijk} < \hat{P}$. This would

[29] Ferri, Hernández-Orallo and Modroiu (2009) and Hernández-Orallo, Flach and Ferri (2012) provide a comprehensive analysis of classification evaluation metrics and differentiate between three classes of evaluation metrics. (1) Metrics based on a threshold that provide an error rate on actual classifications as opposed to predicted probabilities. (2) Metrics based on a probabilistic interpretation of error, which capture the difference between the estimated and true probabilities. (3) Metrics based on how the classifier ranks the samples in terms of likelihood to be a purchaser as opposed to a non-purchaser. As mentioned previously, we ignore the first class of metrics since there are idiosyncrasies in how the threshold is set across websites and so do not analyze the actual classifications. We select the most commonly utilized metrics from the latter two classes. From the second class of evaluation metrics we choose the MSE and from the third class we choose the Area Under the ROC Curve (AUC) metric.

[30] For an extended discussion of ROC analysis, see Fawcett (2006).

Figure 7: Sample ROC Curve



Notes: This figure depicts an ROC curve, which maps out the trade-off between type I and type II errors for a classifier as the classification threshold varies. The area under the ROC curve is denoted by AUC and provides a scalar measure of prediction performance.

result in a false positive rate—a rate at which a non-purchaser is misclassified into a purchaser:

$$FPR := \frac{\text{False Positive}}{\text{True Negative} + \text{False Positive}} = \frac{\sum_{ijk} |\{\hat{p}_{ijk} > \hat{P}, y_{ijk} = 0\}|}{\sum_{ijk} |\{y_{ijk} = 0\}|}.$$

At the same time, it would result in a true positive rate—or a rate at which a purchaser is correctly classified as a purchaser:

$$TPR := \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} = \frac{\sum_{ijk} |\{\hat{p}_{ijk} > \hat{P}, y_{ijk} = 1\}|}{\sum_{ijk} |\{y_{ijk} = 1\}|}.$$

The ROC then depicts the level of TPR a prediction machine achieves for each level of FPR it tolerates.

The ROC is obtained by tracing the locus of (FPR, TPR) by varying the classification threshold $\hat{P}$.[31] The slope of the ROC corresponds to the additional *power* (in rate) the prediction gains for an additional unit of type I error (in rate) it tolerates. For a random predictor, this slope would be one, and the ROC will be a 45 degrees line. A better than random predictor would produce an ROC which lies above that 45 degrees line. Figure 7 depicts a typical ROC curve.

[31] For extreme cases, with $\hat{P} = 1$, all consumers are classified as non-purchasers, which yields $(FPR, TPR) = (0, 0)$, and with $\hat{P} = 0$ all consumers are classified as purchasers, which yields $(FPR, TPR) = (1, 1)$.

The AUC, which measures the area under the ROC, provides a simple scalar measure of the prediction performance. If either the prediction technology improves or the consumer becomes more predictable, then the ROC will shift up and AUC will increase. Aside from the fact that the intermediary focuses on this measure, the AUC is invariant to the change in class distribution (Fawcett, 2006). Suppose for instance the proportion of purchasers increases, as long as the prediction technology remains unchanged, this does not alter FPR and TPR, so the ROC and AUC remain unchanged.

These two measures capture different aspects: AUC captures the ability for the classifier to separate the two different classes whereas MSE captures the accuracy of the estimated probabilities. Hence, we will report the effect on both since they provide two qualitatively different measures of prediction performance.

6.2 Prediction Performance

In this section we investigate the impact of GDPR on predictability at the immediate onset of its implementation. We utilize the same empirical strategy that we described in section 3. The same empirical design is valid because the intermediary trains separate models for each website using only the data from the respective website. As a result, any changes to the collected data from EU websites due to GDPR should not impact non-EU websites. However, there are two limiting factors in our analysis. The first is the restriction on the data; unlike the search and advertising data, the prediction performance requires additional purchase data, which is limited for only a subset of websites.[32] The second is that the models are trained utilizing a sliding window of the data, which means that, even if there is a sudden change to the underlying data distribution, there may be a slow adjustment period that would vary across the different websites.

Table 4 displays the difference-in-differences estimates for all of the relevant prediction related outcome variables. First, column (1) shows that GDPR results in a small but significant increase in the proportion of purchasers. Meanwhile, the insignificant coefficient for average predictive probability in column (2) shows that little adjustment by the classifier of the firm to this change. Figure 14 displays the time-varying specification for these outcome variables indicating that the average predicted probability remains constant whereas the class proportion fluctuates but appears to increase.

[32]We drop websites that either have no purchase data or where the class proportion is degenerate so that $y_{ijk} = 0$ or $y_{ijk} = 1$ for all cookies i on website j. There are also two websites that we know had a reporting error for purchase data during our sample period and we drop them from our analysis. Further, we drop any $(browser, OS, product, website, country)$ tuple that, on average, has fewer than 50 consumers a week since these observations are very noisy due to low sample sizes and the performance of the prediction problem is less interesting in these cases.

Table 4: Difference-in-Differences Estimates for Prediction Outcome Variables

	(1) Class Proportion	(2) Average Predicted Probability	(3) MSE	(4) AUC	(5) Purchaser MSE	(6) Non-Purchaser MSE
DiD Coefficient	0.00915*	0.00129	0.0130***	0.0124	-0.00579	-0.00126
	(1.77)	(0.17)	(3.74)	(1.12)	(-0.43)	(-0.45)
Product Type Controls	✓	✓	✓	✓	✓	✓
OS + Browser Controls	✓	✓	✓	✓	✓	✓
Week FE	✓	✓	✓	✓	✓	✓
Website × Country FE	✓	✓	✓	✓	✓	✓
Observations	15470	15470	15470	15470	14298	15470

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: t-statistics are reported in parentheses. The standard errors for every regression are clustered at the website-country level. We aggregate every dependent variable to the website-country-product type-week level between weeks 16 and 29, including both weeks 16 and 29 (April 13th - July 20th). The dependent variable in the regression reported in the first column is the proportion of purchasers associated with each observation and the second column is the average predicted probability. The dependent variables in the third and fourth column are the MSE and AUC, respectively. Finally, in the fifth and sixth columns the dependent variables are the MSE conditional on the true class of the observation.

Columns (3) and (4) show the impact of GDPR on the prediction performance of the intermediary as measured in MSE and AUC, respectively. Column (3) shows a significant increase in MSE after GDPR. However, rather than indicating the worsened prediction performance, this is likely to be an artifact of the change in class proportion and the lack of adjustment by the classifier.[33] Indeed, columns (5) and (6) show that MSE conditional on true class has not gone up; if anything, they have gone down albeit statistically insignificantly. As mentioned above, given the skewed distribution, an increase in the proportion of purchasers will raise the MSE. In fact, column (4) shows a positive estimate for the treatment effect on AUC indicating a marginal improvement in prediction, though it is not statistically significant. The marginal improvement in AUC indicates that, even in spite of the decreased accuracy of the estimated probabilities, the ability of the intermediary to separate the two classes has increased. This observation is consistent with what we would expect from the aforementioned hypothesis of privacy means substitution.

Finally, Figure 15 displays the results from the time-varying specification for MSE and AUC, indicating that there was an initial increase in MSE followed by an eventual decline. This is consistent with the claim that much of the increase in MSE was a result of the lack of rapid

[33] Appendix F decomposes the change of MSE to accounts for the extent to which the increase may have resulted from the classifier's lack of rapid adjustment to the post-GDPR consumer distribution leading the estimated class probabilities to no longer as closely match the empirical class probabilities.

adjustment. Further, the increases in AUC do not occur directly after GDPR but rather also occur gradually.

Overall, our results suggest that GDPR has not negatively impacted the ability to predict consumer behavior and if at all, the sign of the treatment effect suggests the opposite. This is further validated by the exercise in Appendix G which identifies the expected "long run" changes in prediction performance as a result of the changes to the data observed in section 4. This exercise shows that the increase in trackability ought to lead to improvements to prediction performance, whereas the change in the overall size of data as a result of GDPR should not adversely impact prediction performance significantly.

7 Conclusion

In this paper we empirically study the effects of data privacy regulation by exploiting the introduction of GDPR as a natural experiment. We use data from an intermediary that contracts with many online travel agencies worldwide, which allows us to investigate the effect of GDPR on a comprehensive set of outcomes. Our analysis focuses on the stipulation of GDPR that requires firms to ask consumers for explicit consent to store and process their data.

Our results paint a novel and interesting picture of how a consumer's privacy decision— particularly the means by which she protects her privacy—may impact the rest of the economy, including other consumers, and the firms and advertisers relying on consumer data. The strong and effective means of privacy protection made available by laws such as GDPR and the recent CCPA (California Consumer Privacy Act) should help the privacy-concerned consumers to protect their privacy by eliminating their digital footprints. These consumers are thus clear winners of the laws. However, the impacts on the others are less clear. Our results suggest the possibility that a consumer's switching of the means of privacy protection makes the opt-in consumers who share their data more trackable and possibly more predictable to the firms with which they share data. If this increased identifiability makes up for decreased data (resulting from opt-outs), as indicated by Appendix G, then the firms using consumer data could also come out as winners. What about those consumers who opt in? Their welfare will depend on how their data is used by the firms. If their data is used to target advertising and services to their needs, they too could very well be winners of privacy laws, even if their decision to opt in may not have accounted for the externality. However, if their data is used for extracting consumer surplus, e.g., via personalized pricing, the externalities could harm them.

While these qualitative implications are clear, our reduced-form approach does not allow us to quantify the welfare implications for both consumers and advertisers. We leave for future work a structural analysis of the interactions that we identify in order to better understand the mag-

nitude of each of the channels by which consumers and advertisers are affected. Given the large compliance costs associated with data privacy regulation, decomposing the welfare effects in this manner is a fruitful direction for research and important for further building on our insights in order to guide the design and understanding the value of such regulation.

Finally, our paper has broader implications beyond the online travel industry and keyword-based advertising markets. Firms in this industry, as with many markets in the digital economy, increasingly compete with the large technology firms such as Google whose reach expands across many different online markets and for whom consumers have little choice but to accept data processing. As a result, while our results highlight that increased consent requirements may not be wholly negative for firms, if consumers are similarly using such opt-out capabilities at our estimated rates in other markets (such as behaviorally-targeted advertising markets) then another important direction for future work is understanding the extent to which such regulation puts firms in these markets at a disadvantage relative to these larger firms. We believe that these insights and directions for future work are useful for the design of the many proposed regulations in the US and around the world that follow in the footsteps of GDPR.

References

Abowd, John M, and Ian M Schmutte. 2019. "An economic analysis of privacy protection and statistical accuracy as social choices." *American Economic Review*, 109(1): 171–202.

Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar. 2019. "Too Much Data: Prices and Inefficiencies in Data Markets." National Bureau of Economic Research.

Acquisti, Alessandro, Curtis Taylor, and Liad Wagman. 2016. "The economics of privacy." *Journal of Economic Literature*, 54(2): 442–92.

Acquisti, Alessandro, Leslie K John, and George Loewenstein. 2013. "What is privacy worth?" *The Journal of Legal Studies*, 42(2): 249–274.

Aridor, Guy, Kevin Liu, Aleksandrs Slivkins, and Zhiwei Steven Wu. 2019. "The perils of exploration under competition: A computational modeling approach." 171–172.

Athey, Susan, Christian Catalini, and Catherine Tucker. 2017. "The digital privacy paradox: Small money, small costs, small talk." National Bureau of Economic Research.

Bajari, Patrick, Victor Chernozhukov, Ali Hortaçsu, and Junichi Suzuki. 2019. "The impact of big data on firm performance: An empirical investigation." Vol. 109, 33–37.

Bellemare, Marc F, and Casey J Wichman. 2019. "Elasticities and the inverse hyperbolic sine transformation." *Oxford Bulletin of Economics and Statistics*.

Berendt, Bettina, Oliver Günther, and Sarah Spiekermann. 2005. "Privacy in e-commerce: stated preferences vs. actual behavior." *Communications of the ACM*, 48(4): 101–106.

Bergemann, Dirk, Alessandro Bonatti, and Tan Gan. 2019. "The Economics of Social Data."

Boerman, Sophie C, Sanne Kruikemeier, and Frederik J Zuiderveen Borgesius. 2018. "Exploring motivations for online privacy protection behavior: Insights from panel data." *Communication Research*, 0093650218800915.

Braghieri, Luca. 2019. "Targeted advertising and price discrimination in intermediated online markets." *Available at SSRN 3072692*.

Campbell, James, Avi Goldfarb, and Catherine Tucker. 2015. "Privacy regulation and market structure." *Journal of Economics & Management Strategy*, 24(1): 47–73.

Chiou, Lesley, and Catherine Tucker. 2017. "Search engines and data retention: Implications for privacy and antitrust." National Bureau of Economic Research.

Choi, Jay Pil, Doh-Shin Jeon, and Byung-Cheol Kim. 2019. "Privacy and personal data collection with information externalities." *Journal of Public Economics*, 173: 113–124.

Decarolis, Francesco, and Gabriele Rovigatti. 2019. "From Mad Men to Maths Men: Concentration and Buyer Power in Online Advertising."

Degeling, Martin, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub,

and Thorsten Holz. 2018. "We Value Your Privacy… Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy." *arXiv preprint arXiv:1808.05096*.

DeGroot, Morris H, and Stephen E Fienberg. 1983. "The comparison and evaluation of forecasters." *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2): 12–22.

Dwork, Cynthia. 2011. "Differential privacy." *Encyclopedia of Cryptography and Security*, 338–340.

Dwork, Cynthia, Aaron Roth, et al. 2014. "The algorithmic foundations of differential privacy." *Foundations and Trends® in Theoretical Computer Science*, 9(3–4): 211–407.

Fairfield, Joshua AT, and Christoph Engel. 2015. "Privacy as a public good." *Duke LJ*, 65: 385.

Fawcett, Tom. 2006. "An introduction to ROC analysis." *Pattern recognition letters*, 27(8): 861–874.

Ferri, César, José Hernández-Orallo, and R Modroiu. 2009. "An experimental comparison of performance measures for classification." *Pattern Recognition Letters*, 30(1): 27–38.

Goldberg, Samuel, Garrett Johnson, and Scott Shriver. 2019. "Regulating Privacy Online: The Early Impact of the GDPR on European Web Traffic & E-Commerce Outcomes." *Available at SSRN 3421731*.

Goldfarb, Avi, and Catherine E Tucker. 2011. "Privacy regulation and online advertising." *Management science*, 57(1): 57–71.

Goldfarb, Avi, and Catherine Tucker. 2012*a*. "Privacy and innovation." *Innovation policy and the economy*, 12(1): 65–90.

Goldfarb, Avi, and Catherine Tucker. 2012*b*. "Shifts in privacy concerns." *American Economic Review*, 102(3): 349–53.

Hernández-Orallo, José, Peter Flach, and Cèsar Ferri. 2012. "A unified view of performance metrics: translating threshold choice into expected classification loss." *Journal of Machine Learning Research*, 13(Oct): 2813–2869.

Jia, Jian, Ginger Zhe Jin, and Liad Wagman. 2018. "The short-run effects of GDPR on technology venture investment." National Bureau of Economic Research.

Johnson, Garrett. 2013. "The impact of privacy policy on the auction market for online display advertising."

Johnson, Garrett, and Scott Shriver. 2019. "Privacy & market concentration: Intended & unintended consequences of the GDPR." *Available at SSRN*.

Johnson, Garrett A, Scott K Shriver, and Shaoyin Du. 2020. "Consumer privacy choice in online advertising: Who opts out and at what cost to industry?" *Marketing Science*.

Kehoe, Patrick J, Bradley J Larsen, and Elena Pastorino. 2018. "Dynamic Competition in the Era of Big Data." Working paper, Stanford University and Federal Reserve Bank of Minneapolis.

Liang, Annie, and Erik Madsen. 2019. "Data Sharing and Incentives." *Available at SSRN 3485776*.

Lin, Tesary. 2019. "Valuing Intrinsic and Instrumental Preferences for Privacy." *Available at SSRN*

3406412.

Norberg, Patricia A, Daniel R Horne, and David A Horne. 2007. "The privacy paradox: Personal information disclosure intentions versus behaviors." *Journal of consumer affairs*, 41(1): 100–126.

Utz, Christine, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. 2019. "(Un) informed Consent: Studying GDPR Consent Notices in the Field." 973–990, ACM.

Zhuo, Ran, Bradley Huffaker, Shane Greenstein, et al. 2019. "The Impact of the General Data Protection Regulation on Internet Interconnection." National Bureau of Economic Research.

Appendix

A Additional Consumer Response Figures

Figure 8: Week by Week Treatment Effect (Cookies and Recorded Searches)

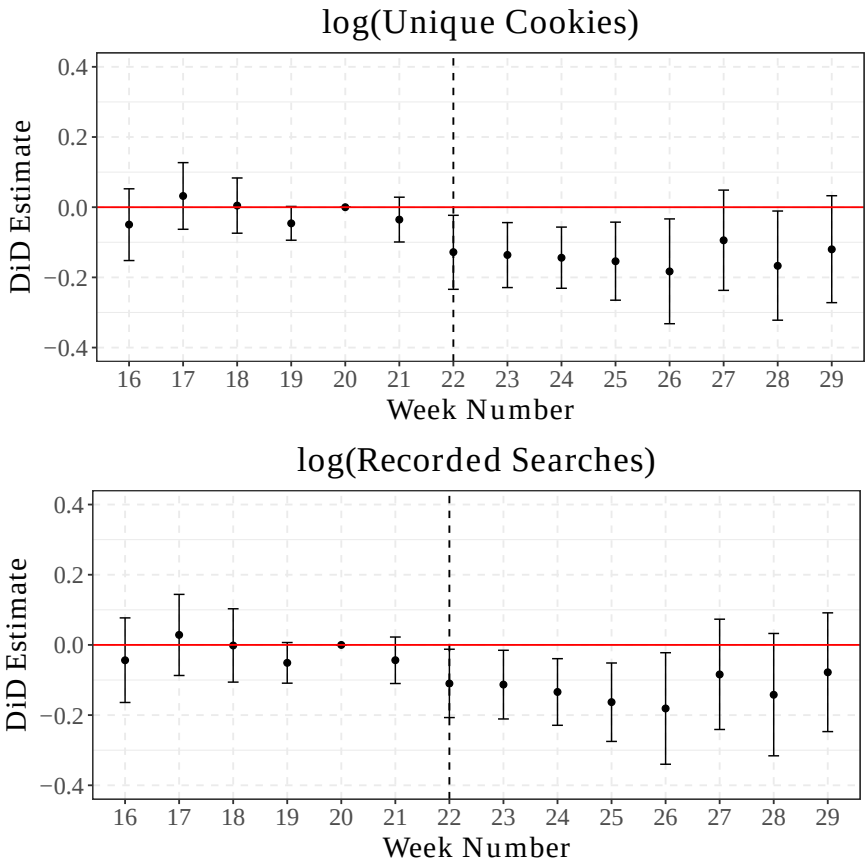


Table 5: Difference-in-Differences Estimates for Sales Activity

	(1) Total Pages	(2) Total Advertising Units
DiD Coefficient	-0.0387	0.0837
	(-0.58)	(1.11)
Product Category Controls	✓	✓
Week FE	✓	✓
Website × Country FE	✓	✓
Observations	3731	3731

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: t-statistics are reported in parentheses. The standard errors for every regression are clustered at the website-country level. We aggregate every dependent variable to the website-country-product type-week level between weeks 16 and 29, including both weeks 16 and 29 (April 13th - July 20th). The dependent variable in the first regression is the total number of pages where the intermediary is present. The dependent variable in the second regression is the total number of advertising units associated with the intermediary.

Figure 9: Week by Week Treatment Effect (Consumer Persistence)

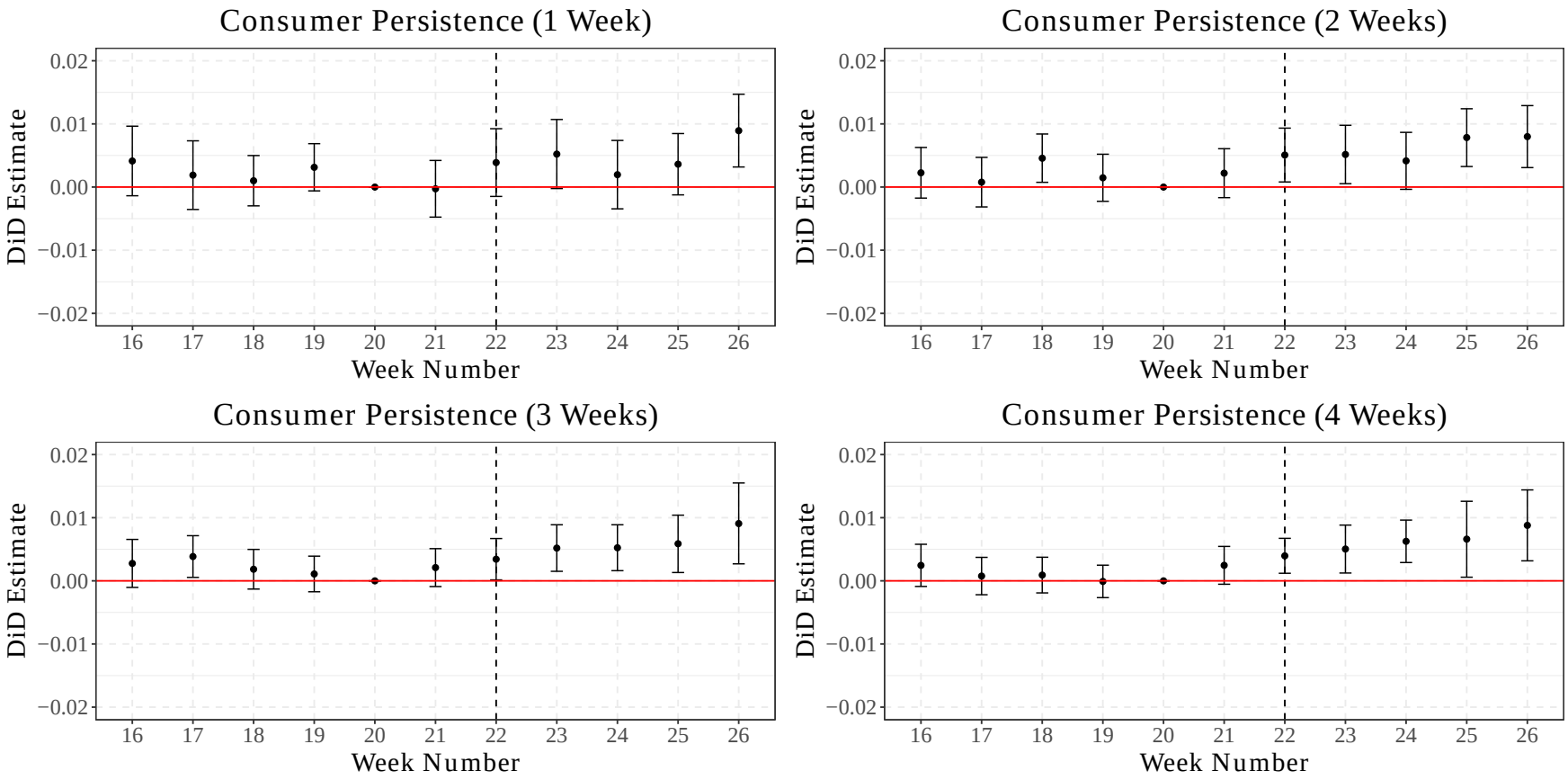
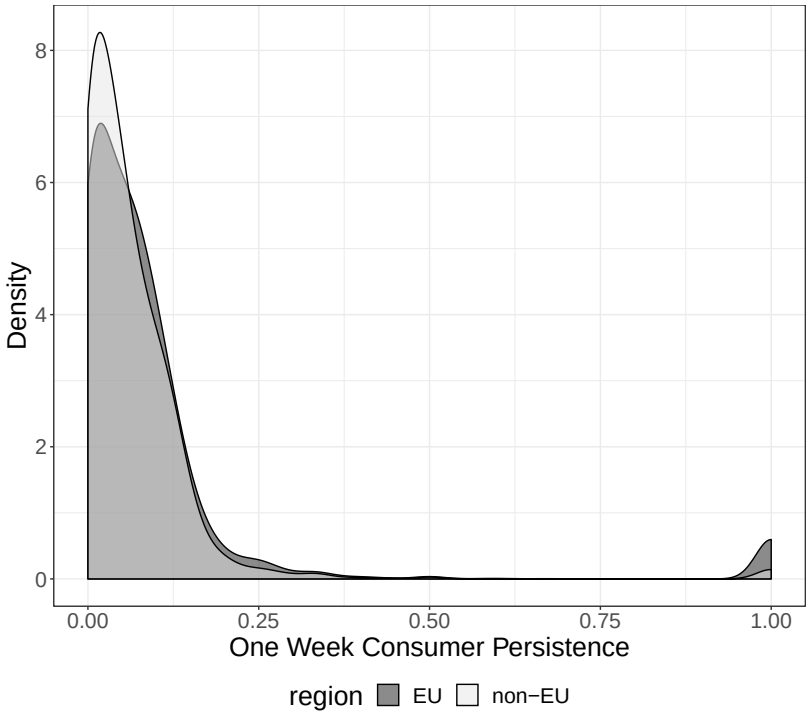


Table 6: Summary Statistics of Consumer Persistence

Treatment Group	1 Week	2 Weeks	3 Weeks	4 Weeks
non-EU	.0640	.0417	.0330	.0282
EU	.0962	.0730	.0644	.0597

Notes: The summary statistics are computed on the sample period before GDPR and show the mean consumer persistence values across the EU and the non-EU for $k = 1, 2, 3, 4$.

Figure 10: Distribution of Consumer Persistence (1 Week)



B Consumer Persistence Heterogeneous Treatment Effects

We further investigate the mechanisms behind the increased consumer persistence by estimating heterogeneous treatment effects across web browsers and operating systems. We exploit the fact that different browsers and operating systems attract different types of individuals with different levels of technical sophistication as well as provide different levels of privacy protection. This exercise provides additional evidence to disentangle the selective consent and privacy means substitution hypotheses since the selective consent hypothesis would predict that there should be no heterogeneity in persistence across these dimensions whereas the privacy means substitution hypothesis would predict the opposite.

First, we study heterogeneous treatment effects across web browsers and restrict attention to the most popular web browsers: Google Chrome, Microsoft Edge, Mozilla Firefox, Internet Explorer, Opera, and Apple Safari. We consider the following specification:

$$y_{tcjobp} = \alpha_t + \delta_{jc} + \kappa_c + \xi_j + \gamma_o + \zeta_b + \omega_p + \beta(EU_j \times after \times browser) + \epsilon_{tcjobp} \tag{5}$$

There are two dimensions on which we could think that the differential change in persistence would vary across web browsers. The first is that there is a demographic selection into browsers and the ability to substitute between various privacy means requires technical sophistication (i.e. consumers need to know how to manage cookies). For instance, Internet Explorer (IE) is a web browser primarily used on older computers and is known to attract older, less technologically sophisticated, users. Thus, the privacy means substitution hypothesis seems more plausible if the effects are stronger on browsers with more technologically sophisticated consumers. The second is that there is different levels of privacy protection among browsers. For instance, Apple Safari at the time of the GDPR had a broad set of privacy protection means built into it, whereas Google Chrome had laxer privacy controls. As a result, we might expect that Safari users would value the privacy protection means offered by GDPR less and thus lead to a smaller increase in persistence.

Table 8 displays the regression results for this specification with Chrome as the omitted browser. The treatment effect is consistent across browsers with the exception of Internet Explorer which has almost no change in persistence. The estimated treatment effect is lower in Safari relative to Chrome, but not significantly so. This provides further evidence for the privacy means substitution hypothesis.

Next, we study heterogeneous treatment effects across operating systems and narrow down the sample to only look at the most popular operating systems: Android, Chrome OS, iOS, Linux, Mac OS X, and Windows. We consider the following specification:

$$y_{tcjobp} = \alpha_t + \delta_{jc} + \kappa_c + \xi_j + \gamma_o + \zeta_b + \omega_p + \beta(EU_j \times after \times OS) + \epsilon_{tcjobp} \tag{6}$$

An important distinction is between mobile operating systems and desktop operating systems. There are less readily available privacy means for cookie management on the mobile web compared to desktop and consumer behavior in general tends to be different on mobile compared to desktop. For consistency with the privacy means substitution hypothesis, we would expect a larger difference in persistence on desktop compared to mobile whereas for consistency with the selective consent hypothesis we should expect no differences.

Table 7 displays the regression results with Windows as the omitted operating system that indicates that Android and iOS have no or weak increases in persistence for $k = 1, 2$ but appear to have an increase in persistence for $k = 3, 4$. This effect is significant and strongest for Android. Otherwise, the treatment effect is approximately the same across the different operating systems. Since there seems to be a weak difference between persistence on mobile and desktop this appears to favor the privacy means substitution effect, but does not provide conclusive evidence.

Table 7: Consumer Persistence by Week - OS Heterogeneous Treatment Effects

	(1) 1 Week	(2) 2 Weeks	(3) 3 Weeks	(4) 4 Weeks
Treated	0.00603***	0.00462***	0.00460***	0.00476***
	(2.70)	(2.76)	(2.65)	(2.91)
Treated × (OS =ANDROID)	-0.00886***	-0.00429*	-0.00256	0.000311
	(-3.19)	(-1.96)	(-1.26)	(0.17)
Treated × (OS = CHROME_OS)	-0.00384	-0.00592	-0.00593	0.00176
	(-0.67)	(-1.24)	(-1.44)	(0.52)
Treated × (OS =iOS)	-0.00367	-0.00184	0.000438	0.00132
	(-1.29)	(-0.77)	(0.19)	(0.70)
Treated × (OS = LINUX)	-0.000856	0.00326	-0.000188	0.000463
	(-0.18)	(0.77)	(-0.06)	(0.12)
Treated × (OS = MAC_OS_X)	-0.00291	-0.000367	-0.00209	-0.00184
	(-1.08)	(-0.19)	(-1.26)	(-1.10)
OS = ANDROID	0.0105***	0.00565**	0.00335	0.00296
	(3.56)	(2.01)	(1.20)	(1.18)
OS = CHROME_OS	0.00307	0.00221	-0.000749	-0.00117
	(0.89)	(0.59)	(-0.27)	(-0.45)
OS = iOS	0.00712***	0.000500	-0.0000303	-0.0000989
	(2.66)	(0.22)	(-0.01)	(-0.05)
OS = LINUX	-0.0164***	-0.0119***	-0.0105***	-0.00732***
	(-4.37)	(-3.46)	(-4.17)	(-2.87)
OS = MAC_OS_X	-0.000548	-0.00115	-0.00299*	-0.00297***
	(-0.24)	(-0.58)	(-1.96)	(-2.68)
Constant	0.0835***	0.0619***	0.0557***	0.0497***
	(33.88)	(29.13)	(31.75)	(29.66)
Product Type Controls	✓	✓	✓	✓
OS × Week, OS × EU Controls	✓	✓	✓	✓
Week FE	✓	✓	✓	✓
Website × Country FE	✓	✓	✓	✓
Browser Controls	✓	✓	✓	✓
Observations	48301	48301	48301	48301

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: t-statistics are reported in parentheses. The standard errors for every regression are clustered at the website-country level. We aggregate every dependent variable to the website-country-product type-week level between weeks 16 and 26, including both weeks 16 and 26 (April 13th - June 29th). We restrict focus only to the most popular operating systems. The dependent variables in the regression are the consumer persistence measures for $k = 1, 2, 3, 4$, respectively. $treated$ indicates whether the observation is associated with an EU website and past the GDPR implementation date. $treated \times os$ indicates the heterogeneous treatment effect for the specified os. The coefficients on os indicate the estimated values for the os fixed effect. The held-out operating system is Windows.

Table 8: Consumer Persistence - Browser Heterogeneous Treatment Effects

	(1) 1 Week	(2) 2 Weeks	(3) 3 Weeks	(4) 4 Weeks
Treated	0.00615***	0.00645***	0.00519***	0.00628***
	(2.99)	(3.51)	(3.29)	(3.49)
Treated × (Browser = EDGE)	-0.00134	-0.00169	0.00230	0.000132
	(-0.35)	(-0.61)	(0.74)	(0.04)
Treated × (Browser = FIREFOX)	-0.00413	-0.00214	-0.00260	-0.00166
	(-1.60)	(-0.89)	(-1.43)	(-0.84)
Treated × (Browser = IE)	-0.0101**	-0.00838***	-0.00375	-0.00497**
	(-2.53)	(-2.67)	(-1.54)	(-2.03)
Treated × (Browser = OPERA)	-0.00935*	-0.00396	-0.00344	-0.00335
	(-1.95)	(-0.83)	(-0.94)	(-0.86)
Treated × (Browser = SAFARI)	-0.00185	-0.00332	-0.00280	-0.00225
	(-0.69)	(-1.43)	(-1.44)	(-1.12)
Browser = EDGE	0.00125	-0.00226	-0.00144	-0.000568
	(0.36)	(-0.78)	(-0.42)	(-0.18)
Browser = FIREFOX	-0.00503**	-0.00381*	-0.00465***	-0.00409***
	(-2.29)	(-1.96)	(-3.13)	(-2.92)
Browser = IE	-0.0164***	-0.0113***	-0.00801***	-0.00764***
	(-6.73)	(-5.15)	(-3.29)	(-4.18)
Browser = OPERA	-0.00151	-0.00337	-0.00665**	-0.00596**
	(-0.39)	(-1.00)	(-2.22)	(-2.15)
Browser = SAFARI	-0.00315	-0.00229	-0.00309*	-0.00211
	(-1.22)	(-1.06)	(-1.80)	(-1.20)
Constant	0.0861***	0.0647***	0.0575***	0.0568***
	(32.30)	(29.30)	(34.48)	(12.53)
Product Type Controls	✓	✓	✓	✓
OS × Week, OS × EU Controls	✓	✓	✓	✓
Week FE	✓	✓	✓	✓
Website × Country FE	✓	✓	✓	✓
OS Controls	✓	✓	✓	✓
Observations	40810	40810	40810	40810

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: t-statistics are reported in parentheses. The standard errors for every regression are clustered at the website-country level. We aggregate every dependent variable to the website-country-product type-week level between weeks 16 and 26, including both weeks 16 and 26 (April 13th - June 29th). We restrict focus only to the most popular web browsers. The dependent variables in the regression are the consumer persistence measures for $k = 1, 2, 3, 4$, respectively. $treated$ indicates whether the observation is associated with an EU website and past the GDPR implementation date. $treated \times browser$ indicates the heterogeneous treatment effect for the specified $browser$. The coefficients on $browser$ indicate the estimated values for the $browser$ fixed effect. The held-out browser is Google Chrome.

C Advertisement and Auction Figures

Table 9: Difference-in-Differences Estimates for Advertisements Delivered

	(1) Total Advertisements Delivered	(2) asinh(Total Advertisements Delivered)
DiD Coefficient	-2627.2	-0.145
	(-1.61)	(-1.52)
OS + Browser Controls	✓	✓
Product Category Controls	✓	✓
Week FE	✓	✓
Website × Country FE	✓	✓
Observations	62328	62328

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: t-statistics are reported in parentheses. The standard errors for every regression are clustered at the website-country level. We aggregate every dependent variable to the website-country-product type-week level between weeks 16 and 29, including both weeks 16 and 29 (April 13th - July 20th). The dependent variables are the log and overall level of total advertisements delivered to consumers.

Figure 11: Week by Week Treatment Effect (Total Advertisements Delivered)

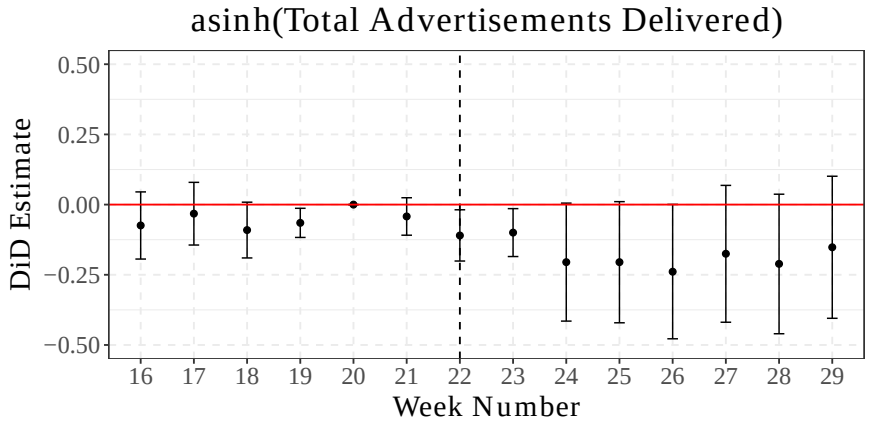


Figure 12: Week by Week Treatment Effect (Average Bid)

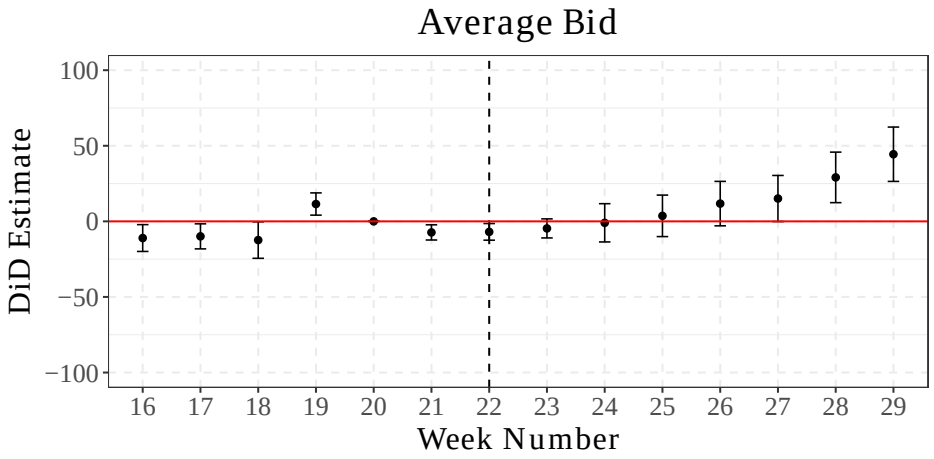


Table 10: Summary Statistics, Bids and Transfers

Treatment Group	Average Bid	Average Transfer
non-EU	394.053	539.652
EU	126.947	174.990

Notes: The table reports the means of the averages bids and average transfer across observations in the pre-GDPR time period for the EU and non-EU respectively.

D Advertisement and Bidder Concentration

In this section we explore whether GDPR resulted in any changes in market concentration among advertisers. We utilize the standard definition for the concentration ratio and the HHI. $CR\text{-}I_{jt}$ is the concentration of impressions for the top I out of K advertisers on website j at time t. Let imp_{kjt} be the impressions of the k-th largest advertiser (according to impression share) on website j at time t. Then, $CR\text{-}I_{jt} = \frac{\sum_{k=1}^{I} imp_{kjt}}{\sum_{k=1}^{K} imp_{kjt}}$. For a website j and time t, the share of advertiser $k \in \{1, 2, ..., K\}$ is denoted by s_k. HHI is therefore defined as follows: $HHI_{jt} = \sum_{k=1}^{K} s_k^2$.

It is important to note that these measures rule out some demand driven changes but not all of them. In particular, advertisers could have churned without any changes in overall concentration and our exercise would not pick this up. However, this is the most plausible and easily measurable channel by which there could be demand-driven changes and so is our primary focus in determining whether the observed results are demand-driven. Table 12 shows the results of

our main specification for CR-1, CR-3, CR-5, and HHI as the outcome variables. There is a statistically significant increase in market concentration across each of these measures, though the effect size is not economically significant. Table 11 displays summary statistics for the various concentration measures and implies that the effect size, for instance, of the increase in market concentration for CR-5 is roughly 3%.

Figure 13 shows the time-varying treatment effect for the various market concentration measures, which all qualitatively follow the same pattern. The increase in market concentration occurs largely leading up until week 20 and remains relatively constant afterwards. However, the increase in average value of a consumer does not occur until week 25 and is a gradual, as opposed to a sudden, increase. Since the change in market concentration does not appear economically significant and the timing of the increase in concentration does not coincide with the increase in the value of a consumer, this provides evidence that the increasing value of consumers is not driven by a change in the composition of advertisers.

Table 11: Summary Statistics, Market Concentration

Treatment Group	**CR-1**	**CR-3**	**CR-5**	**HHI**
Non-EU	.0744	.216	.340	.0546
EU	.0689	.204	.331	.0542

Notes: The table reports the means of several market concentration measures in the pre-GDPR period for both the EU and the non-EU. The first three columns display the mean market share concentrations of the top 1, top 3, and top 5 advertisers according to share of advertisements delivered to consumers (concentration ratio 1, 3, and 5 respectively). The fourth column displays the mean Herfindahl-Hirschman Index (HHI) using the same market share definition.

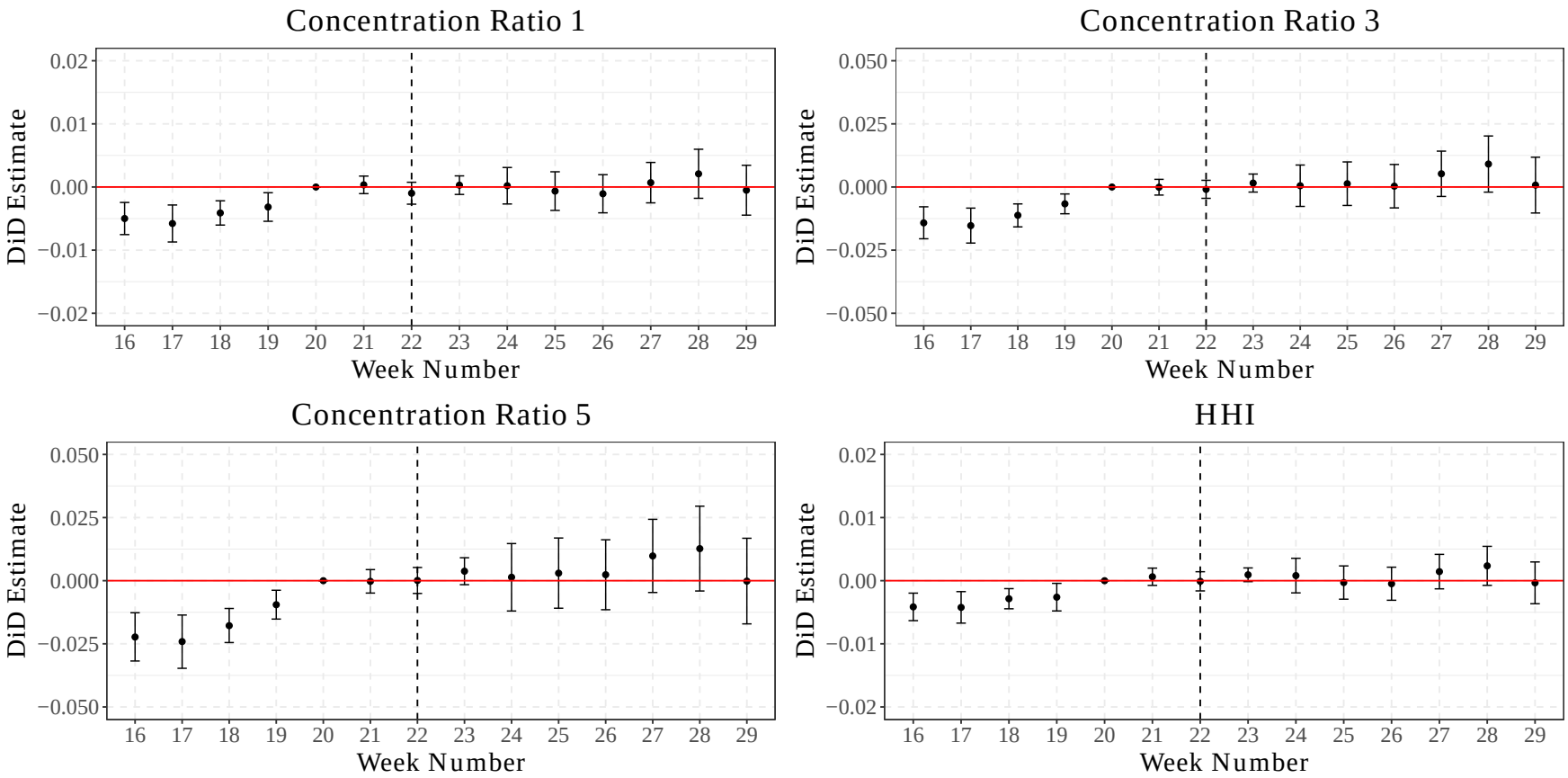
Table 12: Difference-in-Differences Estimates for Market Concentration

	(1) CR 1	(2) CR 3	(3) CR 5	(4) HHI
DiD Coefficient	0.00233*	0.00826**	0.0138**	0.00227*
	(1.78)	(2.17)	(2.24)	(1.95)
OS + Browser Controls	✓	✓	✓	✓
Product Category Controls	✓	✓	✓	✓
Week FE	✓	✓	✓	✓
Website × Country FE	✓	✓	✓	✓
Observations	62328	62328	62328	62328

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: t-statistics are reported in parentheses. The standard errors for every regression are clustered at the website-country level. We aggregate every dependent variable to the website-country-product type-week level between weeks 16 and 29, including both weeks 16 and 29 (April 13th -July 20th). The dependent variables in the regressions reported in the first three columns are the market share concentrations of the top 1, top 3, and top 5 advertisers according to share of advertisements delivered to consumers (concentration ratio 1, 3, and 5 respectively). The dependent variable in the fourth column is the Herfindahl-Hirschman Index (HHI) using the same market share definition.

Figure 13: Week by Week Treatment Effect (Market Concentration)



E Prediction Figures

Figure 14: Week by Week Treatment Effect (Average Predicted Probability and Class Proportion)

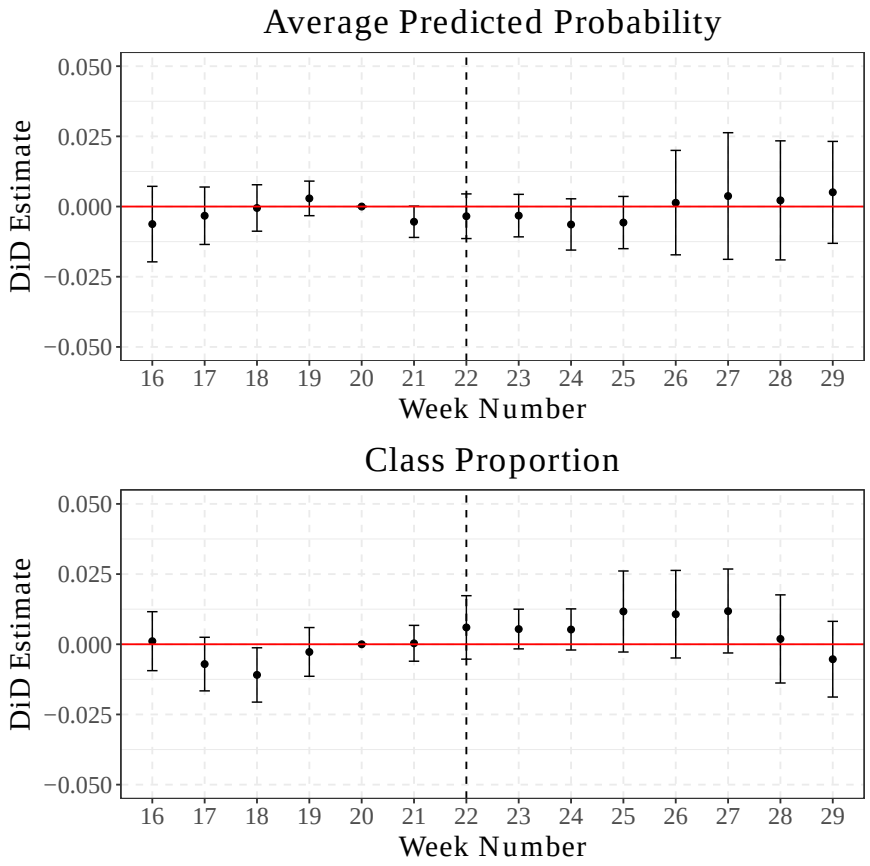
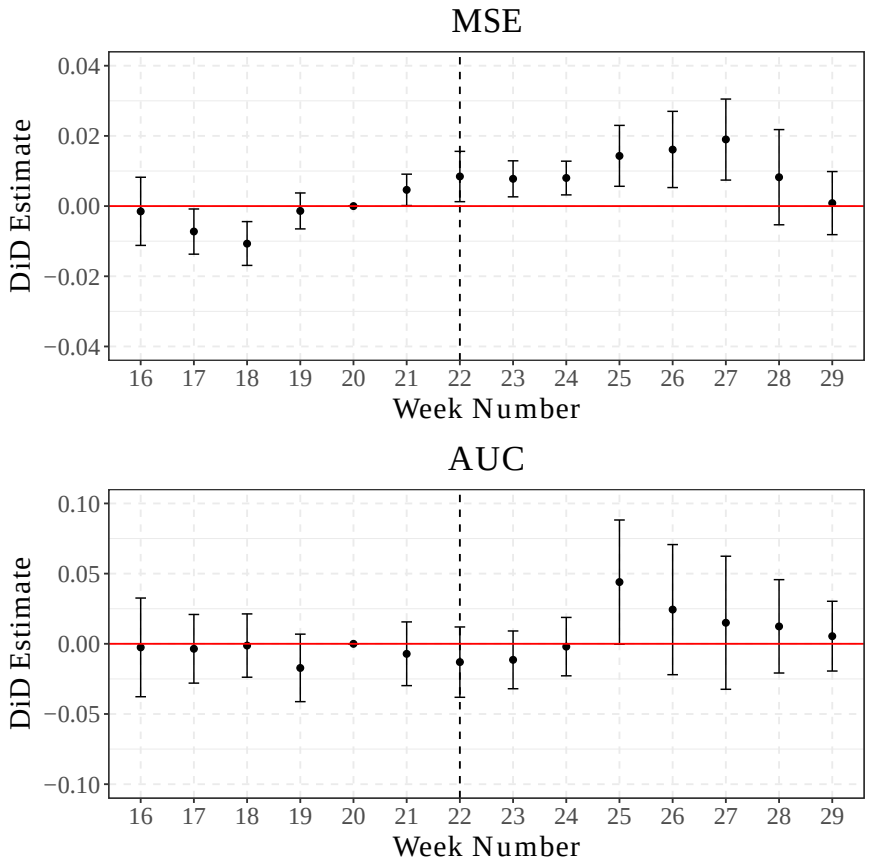


Figure 15: Week by Week Treatment Effect (MSE and AUC)



F Breakdown of MSE

In this section we further investigate the cause of the increase in MSE in our difference-in-differences analysis in section 6. In order to do so we utilize a standard decomposition for the MSE in the classification context and study the effects of GDPR on each component of the decomposition. The MSE for binary classification problems can be decomposed into a *calibration* and *refinement* component (DeGroot and Fienberg, 1983). The *calibration* component indicates the degree to which the estimated probabilities match the true class proportion. The *refinement* component indicates the usefulness of the prediction where a more refined prediction is one that is closer to certainty (i.e. closer to 0 or 1 with 0.5 being the most uncertain). Thus, a classifier with a good MSE is well-calibrated and more refined. This decomposition requires a discretization of the estimated probabilities into a series of K bins.[34] For notation, p_k denotes the kth estimated

[34]Throughout this paper, when calculating the decomposed MSE we will primarily utilize equally spaced bins of size 0.01. Note that since the decomposition requires this discretization, the decomposed MSE and the standard MSE are not precisely the same quantities but are approximately the same.

probability bin, n_k denotes the number of probability estimates falling into the kth bin and $\bar{o}_k$ denotes the true class proportion in the kth bin in the data. This allows us to rewrite (4) as:

$$MSE_j = \underbrace{\frac{1}{\sum_{i \in \mathcal{I}_j} |\mathcal{K}_{ij}|} \sum_{k=1}^{K} n_k (p_k - \bar{o}_k)^2}_{\text{calibration error}} + \underbrace{\frac{1}{\sum_{i \in \mathcal{I}_j} |\mathcal{K}_{ij}|} \sum_{k=1}^{K} n_k \bar{o}_k (1 - \bar{o}_k)}_{\text{refinement error}} \tag{7}$$

We run the same specification utilizing each component of the decomposition of the MSE as the outcome variable. These results are reported in Table 13. They indicate that both the refinement and calibration components increased after GDPR. Both of the components are approximately equally responsible for the increase in MSE with the calibration component being only slightly larger. The increase in calibration error is driven by the classifier's lack of rapid adjustment to the post-GDPR consumer distribution leading the estimated class probabilities to no longer as closely match the empirical class probabilities. However, the increase in refinement error points to a partial adjustment since this increase is a result of the increased uncertainty in the predicted class (i.e. the class proportion moving closer to 0.5.).

Table 13: Difference-in-Differences Estimates for Calibration and Refinement

	(1) Calibration	(2) Refinement
DiD Coefficient	0.00735***	0.00576**
	(2.84)	(2.64)
OS + Browser Controls	✓	✓
Product Category Controls	✓	✓
Week FE	✓	✓
Website × Country FE	✓	✓
Observations	15470	15470

$^{*}\ p < 0.10,\ ^{**}\ p < 0.05,\ ^{***}\ p < 0.01$

Notes: t-statistics are reported in parentheses. The standard errors for every regression are clustered at the website-country level. We aggregate every dependent variable to the website-country-product type-week level between weeks 16 and 29, including both weeks 16 and 29 (April 13th - July 20th). The dependent variable in the regression reported in the first column is the calibration component of the MSE. The dependent variable in the regression reported in the second column is the refinement component of the MSE.

G The Impact of Consumer Persistence and Data Scale on Prediction

The analysis in section 6 on the effect of GDPR on the firm's ability to predict is limited by the data restrictions and the apparent lack of adjustment by its prediction algorithm to the post-GDPR environment. To fully understand the implications for prediction, therefore, we now take a different approach. Instead of asking how the firm's prediction was *actually* impacted in the immediate aftermath, we now ask what would happen to predictive performance in the long run when the algorithm were fully adjusted.

As observed in section 4, GDPR reduces the number of consumers that the intermediary observes but remaining consumers are more persistently trackable. Our approach is to study how these two features—number of observed consumers and the persistence of observed consumers—impact the two measures of prediction performance cross-sectionally by comparing across websites differing in these two dimensions. We use a dataset aggregated at the website-product type-week level. We restrict attention to the pre-GDPR period between January 19th and April 6th. We rely again on the fact that the intermediary only utilizes the data from each individual website in order to train the model for that website. This ensures that predictions for each website are only responsive to the data size and persistence of that website.

We run the following regressions where the dependent variable, $pred_{tcjp}$ represents the prediction error of website j in country c for product type p at time t. The fixed effects are the same as in the primary empirical specification and the standard errors are clustered at the website-country level, the same as with the previous specifications:

$$pred_{tcjp} = \beta \cdot \log(Recorded\_Searches) + \alpha_t + \delta_{jc} + \kappa_c + \xi_j + \omega_p + \epsilon_{tcjobp} \tag{8}$$

$$pred_{tcjp} = \beta \cdot Consumer\_Persistence + \alpha_t + \delta_{jc} + \kappa_c + \xi_j + \omega_p + \epsilon_{tcjobp} \tag{9}$$

Table 14 displays the OLS estimates of the regression relating total recorded searches on prediction error, using both the MSE and AUC as the dependent variables. We report the results of running the regressions with and without the website and website-country fixed effects, but our preferred specification is the one without the website and website-country fixed effects.[35] This corresponds to the regression results in Columns (1) and (3) of Table 14. As expected, an increase in the total recorded searches increases AUC significantly and decreases MSE, albeit insignificantly. Recall that our point estimate of the magnitude of lost data from the GDPR was 10.7%. With this data loss, the magnitude of the predicted decline in prediction error is relatively small

[35] The reason is that the website-country fixed effects soak up the variation in different dataset sizes across websites, even though understanding how this variation impacts prediction error is our main interest.

with a 10.7% decrease in recorded searches only leading to a 0.0007 decrease in AUC.[36]

Table 16 displays the OLS estimates of the regression relating four week consumer persistence to prediction error, using both the MSE and AUC as the dependent variable. As before, we have regressions with and without website and website-country fixed effects, and focus primarily on the regressions without them. Recall that we previously found a 0.00505 increase in the four week persistence as a result of GDPR. Combined with the point estimates from Table 16, this implies an increase of 0.013 for AUC and a decrease of 0.007 for MSE.

Putting these two results together point to the fact that the decline in the overall scale of data should have little impact on predictability, but the change in the nature of the data towards more identifiable consumers should marginally improve prediction according to both AUC and MSE. However, this does not imply that the scale of data is unimportant which would run counter to standard statistical intuition; on the contrary, prediction ability improves substantially as the scale of data increases. Rather, the change in the scale of the data as a result of GDPR is not large enough to cause meaningful changes in prediction error in the long run. However, the increase in persistence as a result of GDPR should lead to an improvement in prediction capabilities in the long run.

[36] In reality the intermediary does not train its models only on data from the current week, but rather utilizing a sliding window of data that includes previous weeks. Table 15 shows the results for the same specification, but uses a sliding window total of recorded searches instead of the weekly total number of recorded searches, and shows that the point estimates do not change much when taking this into account.

Table 14: Prediction Error and Scale of Data

	(1) AUC	(2) AUC	(3) MSE	(4) MSE
log(Recorded Searches)	0.0154*	0.0178	-0.00435	0.000937
	(1.84)	(0.98)	(-0.88)	(0.15)
Constant	0.505***	0.510**	0.191***	0.0987
	(4.60)	(2.45)	(2.82)	(1.31)
Product Category Controls	✓	✓	✓	✓
Week FE	✓	✓	✓	✓
Country FE	✓	✓	✓	✓
Website × Country FE		✓		✓
Observations	874	874	874	874
R^2	0.129	0.699	0.138	0.936

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: t-statistics are reported in parentheses. The standard errors for every regression are clustered at the website-country level. We aggregate every dependent variable to the website-country-product type-week level between weeks 4 and 14, including both weeks 4 and 14 (January 9th - April 5th). The dependent variable in the regression reported in the first and second column is AUC. The dependent variables in the third and fourth column is the MSE. The regression results reported in column (1) and (3) do not include website or website-country fixed effects, whereas those reported in column (2) and (4) include these fixed effects.

Table 15: Sliding Window Data Scale and Aggregate Prediction Error

	(1) AUC	(2) AUC	(3) MSE	(4) MSE
log(Two Week Search Total)	0.0158*		-0.00439	
	(1.88)		(-0.87)	
log(Three Week Search Total)		0.0161*		-0.00440
		(1.92)		(-0.86)
Constant	0.651***	0.479***	0.0942	0.192**
	(5.34)	(4.05)	(1.28)	(2.56)
Product Category Controls	✓	✓	✓	✓
Week FE	✓	✓	✓	✓
Country FE	✓	✓	✓	✓
Website × Country FE		✓		✓
Observations	868	861	868	861
R^2	0.129	0.129	0.140	0.142

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: t-statistics are reported in parentheses. The standard errors for every regression are clustered at the website-country level. We aggregate every dependent variable to the website-country-product type-week level between weeks 4 and 14, including both weeks 4 and 14 (January 9th - April 5th). The dependent variable in the regression reported in the first and second column is AUC. The dependent variables in the third and fourth column is the MSE. The regression results reported in column (1) and (3) do not include website or website-country fixed effects, whereas those reported in column (2) and (4) include these fixed effects. The Two Week Search Total and Three Week Search Total variables are computed by summing the total number of searches observed for each observation over a sliding window of two weeks and three weeks, respectively.

Table 16: Consumer Persistence and Prediction Error

	(1) AUC	(2) AUC	(3) MSE	(4) MSE
Four Week Persistence	2.621***	0.758	-1.401**	0.611*
	(4.55)	(0.95)	(-2.58)	(1.67)
Constant	0.542***	0.686***	0.221***	0.0852***
	(11.35)	(20.17)	(4.91)	(5.30)
Product Category Controls	✓	✓	✓	✓
Week FE	✓	✓	✓	✓
Country FE	✓	✓	✓	✓
Website × Country FE		✓		✓
Observations	874	874	874	874
R^2	0.230	0.691	0.223	0.938

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: t-statistics are reported in parentheses. The standard errors for every regression are clustered at the website-country level. We aggregate every dependent variable to the website-country-product type-week level between weeks 4 and 14, including both weeks 4 and 14 (January 9th - April 5th). The dependent variable in the regression reported in the first and second column is AUC. The dependent variables in the third and fourth column is the MSE. The regression results reported in column (1) and (3) do not include website or website-country fixed effects, whereas those reported in column (2) and (4) include these fixed effects.