The Effect of Privacy Regulation on the Data Industry: Empirical Evidence from GDPR[*]

Guy Aridor[†] Yeon-Koo Che[‡] Tobias Salz[§]

April 19, 2021

Abstract

Utilizing a novel dataset from an online travel intermediary, we study the effects of EU's General Data Protection Regulation (GDPR). The opt-in requirement of GDPR resulted in 12.5% drop in the intermediary-observed consumers, but the remaining consumers are trackable for a longer period of time. These findings are consistent with privacy-conscious consumers substituting away from less efficient privacy protection (e.g, cookie deletion) to explicit opt out—a process that would make opt-in consumers more predictable. Consistent with this hypothesis, the average value of the remaining consumers to advertisers has increased, offsetting some of the losses from consumer opt-outs.

Keywords: GDPR, Data Privacy Regulation, E-Commerce
JEL Codes: L00; L50; K20; L81.

[*] This paper was previously circulated as *The Economic Consequences of Data Privacy Regulation: Empirical Evidence from the GDPR.* We would like to thank Daron Acemoglu, Francesco Decarolis, Glenn Ellison, Sara Ellison, Avi Goldfarb, and seminar participants at Columbia, ASSA 2021, the MaCCI/EPoS Virtual IO Seminar, Columbia Data Science Day, FTC PrivacyCon, Yale Law Big Tech & Antitrust Conference, and the ETH Zurich 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] Beyond the immediate economic harm resulting from such data breaches consumers might also value privacy for its own sake.[4] 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.[5] 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* provides a frontline defense of privacy for consumers: 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. Without them, these firms may not be able to learn and predict consumer behavior and target their services and advertising accordingly.

[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] For the different motivations for privacy, see, for instance: Acquisti, Taylor and Wagman (2016). As noted by Lin (2019), consumer privacy preferences contain both an instrumental and non-instrumental component.

[5] 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.

Our investigation focuses on three broad questions. First, *to what extent do consumers exercise the consent right enabled by GDPR?* Anecdotal and survey evidence suggests that consumers value their privacy. Yet, as commentators argue, consumers may not be willing to act on their privacy concerns even at little cost or inconvenience.[6] We do not yet have clear empirical answers on this question, and consumers' decisions in their GDPR opt-out could shed light on their "revealed" value of privacy.

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 does not eliminate their footprints altogether but rather simply generate "spurious" identifiers that are difficult for firms to distinguish from genuine footprints left by consumers who do not adopt them. 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? These questions are particularly important for the competitive landscape of the digital economy. While big technology firms such as Google or Facebook enjoy virtually unlimited access to consumer data based on their extraordinary reach and presence, many third-party companies can only access the data shared by first-party affiliates. A concern is that the playing field of these firms, already threatened by the big tech companies, may be further weakened by the increased consent requirement of data regulation.[7] How such a third-party firm copes with GDPR could

[6] 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.

[7] This concern has been raised in the recent literature (Johnson, Shriver and Goldberg, 2020; Batikas et al., 2020) and the popular press (GDPR Has Been a Boon for Google and Facebook. `https://www.wsj.com/articles/gdpr-has-been-a-boon-for-google-and-facebook-11560789219`. Accessed on June 2nd, 2020) which show that GDPR led to an increase in market concentration of web trackers, favoring those from Google and Facebook. This concern is exacerbated by the scope of these companies that allows them to collect data across many different devices and domains that are not feasible for smaller third-party vendors (All the Ways Google Tracks You—And How to Stop It. `https://www.wired.com/story/google-tracks-you-privacy/`. Accessed on June 2nd, 2020).

provide a valuable clue on how data regulation may influence the competitive playing field of the digital economy.

To answer these questions, we use the data provided by an anonymous intermediary that contracts with many of the largest online travel agencies and travel meta-search engines around the world. The dataset is uniquely suited for the current inquiries in several respects. An integral part of the intermediary's business is to predict consumer behavior. Upon each visit by a consumer at an online travel agency (its first-party affiliate), this firm predicts the likelihood of the consumer buying from the website and places advertisements from alternative travel agencies to consumers it deems unlikely to purchase from the original website.

The data links consumers' behavior 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 his/her cookies cannot be stored, so the consumer is no longer observed by the intermediary. 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 deployed 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. While the latter generates many “bogus” short-lived consumers (as a new ID is assigned to a consumer, thus making her appear as a new user, each time she visits the site), the former—the GDPR opt-out—simply removes these individuals from the data. As a result, those consumers that remain in the data after the implementation of GDPR are more persistently identifiable.

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 from the third party intermediary. 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.

Our results suggest a novel form of externalities that privacy-conscious consumers exert on the rest of economy—including other consumers and the firms and advertisers relying on consumer data. Their switching away from inefficient browser-based means of privacy protection to an explicit opt-out (enabled by data privacy regulation) could expose the digital footprints of those who choose not to protect their privacy and make them more predictable. These externalities have potentially important implications. First, third-party firms will suffer from loss of consumers who opt out, but this loss will be mitigated by the increased trackability of those consumers who remain. Indeed, our analysis suggests that the mitigating effect could be important; while we find a negative point estimate on overall advertising revenue, this decrease is not statistically significant. Meanwhile, the welfare effect on the remaining consumers depends on how their data is used by the firms. If their data is used to target advertising and services to their needs, as appears to be so far the case, the externality is largely positive and they will be also better off. However, if the data is used to extract their surplus—a possibility in the future—, they could be harmed by the increased trackability.

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: A closely related study that also explores the short run effect of GDPR is Goldberg, Johnson and Shriver (2021). We see these two studies as complementary in terms of the data scenario and findings. Our study utilizes data at a more disaggregate level but is confined to one industry whereas Goldberg, Johnson and Shriver (2021) have a broad cross-section of different websites and are able to investigate to what extent the effect of the GDPR works through a user acquisition channel. Instead, we are able to look in more detail at cookie lifetime and how GDPR has affected advertising revenues of third party firms.

Several other papers have studied the impact of the GDPR in other domains (Jia, Jin and Wagman, 2018, 2020; Zhuo et al., 2019; Utz et al., 2019; Degeling et al., 2018). Batikas et al. (2020); Johnson, Shriver and Goldberg (2020) show that GDPR increased market concentration amongst web technology services. Goldfarb and Tucker (2011); Johnson, Shriver and Du (2020) study the effectiveness of previous data privacy regulations on online advertising. Godinho de Matos and Adjerid (2019) conduct an experiment with a European telecommunications provider to test how consumers respond to the more stringent opt-in requirements that are mandated by GDPR. Finally, Johnson (2013) estimates a structural model of advertising auctions and shows through counterfactual calculations that advertisement revenue drops substantially more under an opt-in rather than an opt-out policy. We complement these papers by utilizing the scope of our setting to tie each of these pieces together and characterize how they interact with each other and are impacted by data privacy regulation.

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.[8,9] 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 informational 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 externalities 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.[10]

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 com-

[8] There is also an emerging, broadly related, literature that studies implications of a more data-driven economy (Goldfarb and Tucker, 2012*a*; Einav and Levin, 2014; Chiou and Tucker, 2017; Kehoe, Larsen and Pastorino, 2018; Aridor et al., 2019; Bajari et al., 2019)

[9] 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.

[10] Our explanation for these externalities is consistent with work which shows that the inability to link consumers over time may lead to difficulties in measuring experimental interventions (Coey and Bailey, 2016; Lin and Misra, 2020).

ponents. 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.[11] 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*

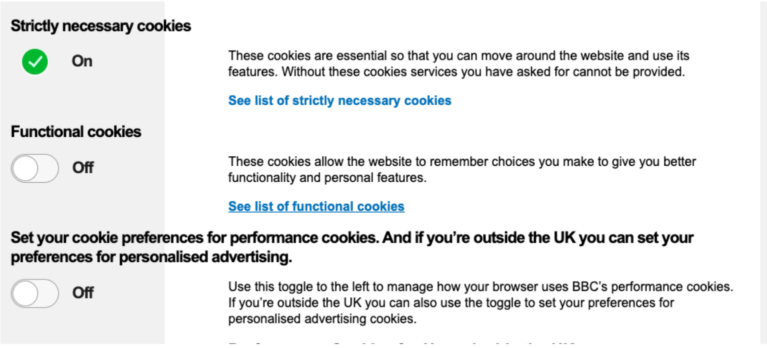
[11] GDPR was intended to overhaul and replace the Data Protection Directive which was enacted in 1995. GDPR further complements the other major European Privacy Regulation, The Privacy and Electronic Communications Directive, also known as the "Cookie Law". Relative to this law, GDPR strengthened the 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.

> *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.

Figure 1: Example Consent Notifications

(a) Post-GDPR consent dialog



Strictly necessary cookies

✓ On

These cookies are essential so that you can move around the website and use its features. Without these cookies services you have asked for cannot be provided.

See list of strictly necessary cookies

Functional cookies

○ Off

These cookies allow the website to remember choices you make to give you better functionality and personal features.

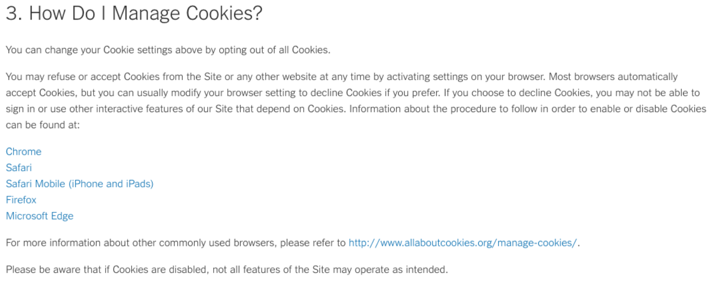
See list of functional cookies

Set your cookie preferences for performance cookies. And if you're outside the UK you can set your preferences for personalised advertising.

○ Off

Use this toggle to the left to manage how your browser uses BBC's performance cookies. If you're outside the UK you can also use the toggle to set your preferences for personalised advertising cookies.

(b) Standard opt-out on US websites



3. How Do I Manage Cookies?

You can change your Cookie settings above by opting out of all Cookies.

You may refuse or accept Cookies from the Site or any other website at any time by activating settings on your browser. Most browsers automatically accept Cookies, but you can usually modify your browser setting to decline Cookies if you prefer. If you choose to decline Cookies, you may not be able to sign in or use other interactive features of our Site that depend on Cookies. Information about the procedure to follow in order to enable or disable Cookies can be found at:

Chrome
Safari
Safari Mobile (iPhone and iPads)
Firefox
Microsoft Edge

For more information about other commonly used browsers, please refer to http://www.allaboutcookies.org/manage-cookies/.

Please be aware that if Cookies are disabled, not all features of the Site may operate as intended.

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.

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 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.[12] 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.

2.2 Consumer-Tracking Technology

The primary consumer tracking method that we focus on in this study are web cookies.[13] 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 manual deletion of cookies, "private browsing" mode,[14] or cookie blockers.[15] 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. One important detail to note is how cookie blockers work in this context. According to our discussions with employees of the intermediary, these services continually regenerate the identifier utilized by the intermediary while still allowing consumers to see the advertisements. Thus, these consumers leave a distinct mark in the data as "single searchers" who only have one observation associated with their identifier.

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

[12] 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.

[13] 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.

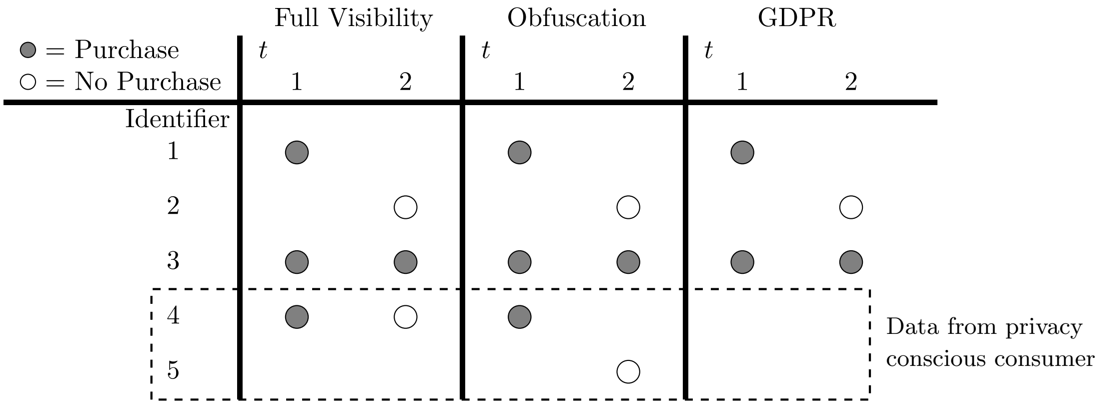
[14] Private browsing modes create "sandbox" browser environments where cookies are only set and used for the duration of the private browsing session. As a result, the website cannot link together data from the same consumer both before and after the private browsing section.

[15] 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)

protection than the aforementioned means since they block all non-essential information from being sent to the third-party website.[16] In our context, this means that by simply opting out consumers can keep their data from being sent to the intermediary since it provides a non-essential, third-party service.

The data generating process, therefore, depends on how consumers protect their privacy. Before GDPR a privacy-conscious consumer would rely on browser-based privacy means, in which case this consumer's data would still be sent to the intermediary but with many "bogus" identifiers associated with the same consumer. By contrast, after GDPR, such a consumer could simply opt out of data sharing, in which case no data on that consumer is sent to the intermediary. This is the important distinction for our purpose. Browser-based privacy means lead to many artificially short consumer histories that still enter the data, whereas GDPR opt-out removes the data completely.[17]

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



	Full Visibility		Obfuscation		GDPR	
	t		*t*		*t*	
	1	2	1	2	1	2
Identifier						
1	●		●		●	
2		○		○		○
3	●	●	●	●	●	●
4	●	○	●			
5				○		

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.

[16] 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.

[17] 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 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.

This is illustrated in Figure 2. The figure shows the data 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 ("Obfuscation") 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 his data completely disappears.

The figure also illustrates how the different data scenarios impact the intermediary's ability to predict consumer behavior, and in particular, how a consumer's choice of privacy means may affect that ability. The four consumers have distinct histories, and these differences may signal different future behavior for them. For example, consumer 4 may be less likely than consumer 1 to purchase from the website next time she visits the website. Under Full Visibility, the prediction machine will correctly recognize this distinction and assign a different prediction score to consumer 4 than to consumer 1. Suppose, however, in the pre-GDPR regime, consumer 4 deletes her cookies and gets partitioned into two separate identifiers, 4 and 5. This behavior confounds the intermediary's ability to predict not only 4's behavior but also 1 and 2's: consumer 1 is now indistinguishable from consumer 4 and consumer 2 is indistinguishable from consumer 5 (the same person as consumer 4) from the intermediary's view point. For instance, the intermediary will assign a lower than accurate purchase odds to consumer 1, influenced by the fact that consumer 4 with the *same* history simply disappears after the visit at $t = 1$. Note that this problem exists even when the intermediary's prediction machine eventually "learns" about the presence of obfuscators, since it cannot tell who obfuscates and who does not. 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 confounding that the intermediary suffered from 4's obfuscation in understanding and predicting 1 and 2's behavior.[18]

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

[18] Importantly, the pre-GDPR intermediary cannot simply replicate the same dataset as post-GDPR since the obfuscators' identities are latent to the intermediary, so their data cannot be surgically cleaned away; for instance, eliminating single-search data will eliminate not only 4 but also 1 and 2 from the data.

agencies (OTAs) in the United States and Europe as well as most prominent travel meta-search engines. We observe consumer searches, online advertising, and the intermediary's prediction of consumer behavior. Our primary analysis utilizes data from this intermediary ranging from April to July 2018.

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.[19] The data was aggregated on a weekly level to remove unimportant day-of-the-week fluctuations. Furthermore, 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.[20]

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

[19] We drop from this aggregation observations which are labeled as coming from bots.

[20] 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 estimates are not biased from entry / exit of websites into our data during the sample period. According to discussions with our data provider, 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.

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 in the time period of interest.

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:

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

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, 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.[21]

In order to validate parallel trends and to understand the persistence of the treatment effect, we further utilize a regression specification that captures the potentially time-varying nature of the treatment:

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

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. Furthermore, week 20 is consistently the baseline week in our regressions since there are some

[21] We cluster at the website-country level because of differences in privacy concerns across countries (Prince and Wallsten, 2020) and differences in consent implementations across websites within jurisdiction (Utz et al., 2019).

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.[22,23]

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. Furthermore, 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.[24]

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.[25] 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.

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.

[22] 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.

[23] Our analysis ends at the end of July since this was the time period over which we were able to obtain data from our data provider.

[24] 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.

[25] 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.

4.1 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.[26]

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

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 $U_{jt}^{OPT-OUT}$.[27,28]

Figure 3 displays the total unique cookies for two multi-national websites, one of which implemented the consent guidelines of the GDPR and the other which does business in the EU but did not immediately comply with the regulations. The multi-national website which implemented the consent guidelines shows a clear drop in observed cookies on European websites 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%. As previously mentioned, our estimates should be interpreted in the context of mixed compliance with the consent guidelines of GDPR as evidenced from Figure 3.

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.

Another measure of consumer response is the total number of searches that are recorded

[26] Note that a website here serves as a first-party affiliate of our intermediary; so the true number of consumers for website j is not the true number of consumers for the intermediary, as opt-out consumers become out of its reach.

[27] As noted in Goldberg, Johnson and Shriver (2021), 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. Furthermore, if a website decided to stop using the intermediary altogether then, as noted previously, they would not be part of our sample.

[28] 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 total advertising units and total pages on which the intermediaries advertising appears as the outcome variables. The results in Table 5 show that there was no significant change in either of those two variables. Thus, we rule this alternative explanation out.

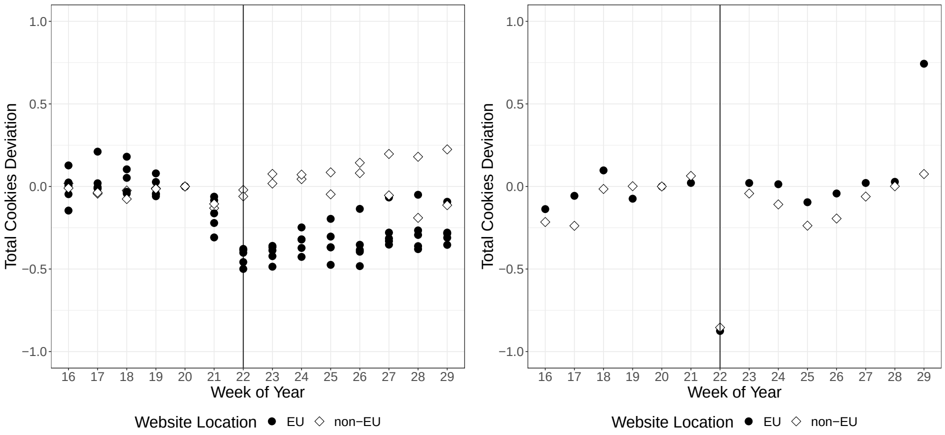
by the intermediary. This outcome measure can also be interpreted as the overall data size observed by the intermediary and how it is affected by GDPR. We re-run the same specification with recorded searches as the dependent variable and report the results in columns (3) and (4) of Table 1. We find that 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 provide evidence for the validity of the difference-in-differences strategy we rely on our time-varying treatment specification. Figure 7 displays the resulting treatment effect over time and points to parallel pre-trends as well as a consistent treatment effect size over our sample period though there is a slight decrease in the estimated treatment effect as we approach the end of our sample period. Finally, as further evidence of robustness, we employ a synthetic control approach, which is reported in subsection B.1 and produces qualitatively similar results.

We want to discuss two further potential threats to the validity of our empirical strategy. The first is a potential contamination between treatment and control groups that may result from multi-national companies implementing the consent mechanisms across all of their websites. The second is that the results may be driven by seasonal travel differences between the treatment and control groups. The first is not a big concern in our setting because multinational online travel agencies serve customers through country-specific websites and have incentives to only make their EU domains compliant with GDPR. For the online travel agencies where we can directly verify compliance we do indeed see that most of them only implement it for their respective EU domains as evidenced by Figure 3. Furthermore, to the extent that there is still residual contamination, it would mean that our estimates are a lower bound of the true effect size.

For the second issue, this is the reason that we focus our analysis on a tight window around the GDPR implementation date and select control countries that ought to have similar travel patterns during this time period. However, since European travel patterns have a somewhat steeper summer gradient than US travel patterns we would expect this to bias against our results. We therefore further supplement our analysis with Google Trends data on travel searches, which should be unaffected by GDPR and provide a good picture into travel trends across these different countries. Using this data, we first graphically show that in the period of the year that we consider, travel patterns between the countries in the analysis are similar. When we augment our primary analysis with country-specific seasonal controls based on Google Trends data we find quantitatively very similar results with slightly stronger effect sizes than before. The full details of this exercise are deferred to subsection B.2.

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



Notes: Each point on the graph represents the total number of unique cookies for a single country, reported in terms of its percent deviation relative to week 20, or $\frac{U_t - U_{t=20}}{U_{t=20}} \quad \forall t \neq 20$. The figure on the left presents a multi-national website that we were able to verify implemented the consent guidelines of GDPR. In this figure, the black dots represent the represented European countries (United Kingdom, France, Germany, Italy, Spain) and the two white dots represent the two non-EU countries where this website functions - the United States and Canada. The figure on the right presents a multi-national website that we were able to verify did not implement the consent guidelines of GDPR. The black dots represent the values from the United Kingdom and the white dots represent the values from the United States.

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-browser-OS-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.

4.2 Persistence of Identifier

A natural question is whether GDPR affects the ability to persistently track consumers. To address this question, 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 $k = 4$ and display the persistence measure for the same two multi-national websites with country-specific versions of their 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 $k \in \{1, 2, 3, 4\}$.[29,30]

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.[31] 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 8 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.[32] 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.

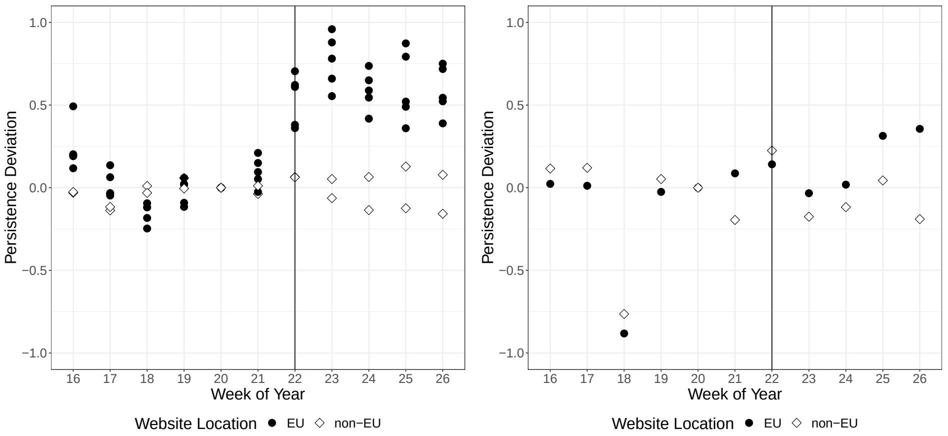
[29] In order to run specification (1) we drop the last 4 weeks of our sample so that we are utilizing the same sample as we vary k. However, our results are qualitatively robust to including these weeks when the data for them is available.

[30] Note that the units on the regression and Figure 4 are not the same. Figure 4 displays the persistence measure in terms of percent deviations from week 20 whereas the coefficients in Table 2 are changes in levels.

[31] 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.

[32] Furthermore, Figure 9 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. They are also robust to the addition of seasonal travel controls using the same procedure as in subsection B.2.

Figure 4: Four Week Persistence for Two Multi-National Websites



Notes: Each point on the graph represents the four week persistence fraction for a single country, reported in terms of its percent deviation relative to week 20, or $\frac{persistence_{4,t} - persistence_{4,t=20}}{persistence_{4,t=20}} \quad \forall t \neq 20$. The figure on the left presents a multi-national website that we were able to verify implemented the consent guidelines of GDPR. In this figure, the black dots represent the represented European countries (United Kingdom, France, Germany, Italy, Spain) and the two white dots represent the two non-EU countries where this website functions - the United States and Canada. The figure on the right presents a multi-national website that we were able to verify did not implement the consent guidelines of GDPR. The black dots represent the values from the United Kingdom and the white dots represent the values from the United States.

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-browser-OS-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.

There are two possible hypotheses for the increased persistence. The first is a *selective consent hypothesis* where consumers only consent to data processing by websites that they frequently use. According to this hypothesis, infrequent users of a website are more likely to opt out of data sharing than frequent users, so the opt-in 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 browser-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.

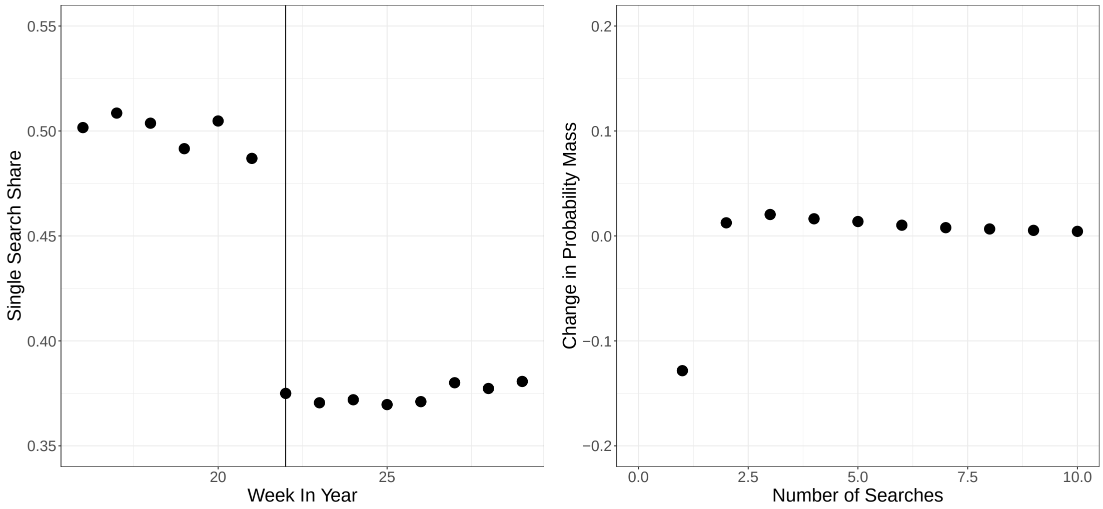
These alternative hypotheses have different economic implications. If the selective consent hypothesis is the predominant explanation for the increased persistence, then privacy regulation may favor firms with more established reputations or offer a wider variety of services.[33] The hypothesis would 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 persistence, then there are several economically relevant consequences. First, the benefit of GDPR would be the marginal benefit over existing privacy protection. Thus, even though a significant fraction of consumers opted out of data collection, the welfare gains for the opted out consumers depends on the marginal privacy gain relative to these pre-existing means. 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 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.

While the two explanations are not mutually exclusive, we provide suggestive evidence that the *privacy means substitution hypothesis* is the more plausible one. Both hypotheses imply that the drop in relative probability mass should be concentrated towards the lower end of the support. However, recall from section 2 that in our context one signature of browser-based privacy protection is a large mass of "single search" consumers. This is due to the continuing regeneration of cookie-identifiers after every request. Indeed, Figure 5 shows that the fraction of single searchers significantly dropped after the implementation of GDPR. Instead, under the selective

[33] 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 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. The figure on the right 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.

consent hypothesis, we would expect that the loss in probability mass would be more evenly distributed across search counts.

Based on this observation, we test for the presence of excess single searchers using a Vuong (1989) test both before and after the implementation of GDPR. The test is based on a simple model in which obfuscating and non-obfuscating consumers together give rise to an observed distribution of cookie counts. We implement this test under two different distributional assumptions. Under the first assumption, the true number of visits per consumer follows a conditional Poisson distribution. To allow for more dispersion we, alternatively, assume that the true underlying distribution follows a negative binomial distribution. For the empirical implementation, we focus on a large website which we know faithfully implemented the consent mechanism. In short, the test suggests the presence of excessive single searchers in the pre-GDPR period but not in the post-GDPR period. The full details of the exercise are deferred to Appendix D.

We implement the test separately for the pre- and post-GDPR period. In the pre-GDPR period, the test rejects a model without excess single searchers in favor of a model in which the number of single searchers is inflated under both distributional assumptions. The same is true for the post-GDPR period under the Poisson distribution, but the estimated fraction of single searchers is lower in this period relative to the pre-GDPR period. However, under the more flexible negative binomial distribution, we still find statistical evidence for excessive single searchers in the pre-

GDPR period whereas the evidence is not significant in the post-GDPR period.

Finally, we analyze the entire set of websites again and estimate heterogeneous treatment effects across popular web browsers and operating systems. We find that the increase in persistence occurs on all browsers except for Internet Explorer and find weak evidence that the increase in persistence is more prominent on desktop operating systems compared to mobile operating systems. While the differences on these dimensions are difficult to explain according to the selective consent hypothesis, they are plausible under the privacy means substitution hypothesis, in light of the alleged lack of technical sophistication by the consumers who use Internet Explorer and the technical difficulty of utilizing browser-based privacy means on mobile/Internet Explorer.[34] The results and a full discussion are deferred to Appendix C. Overall, these results provide additional evidence in favor of the privacy means substitution hypothesis although additional research on this distinction is certainly warranted.

5 GDPR and Online Advertising

The advertisements in our setting are sold via real-time auctions that are held when a consumer makes a search query.[35] Advertisers bid on search keywords such as the origination, destination, or dates of travel. For example, an advertiser may submit a bid to show an advertisement for a consumer searching for a flight from JFK to LAX and upon winning displays a price comparison advertisement for this particular route. Thus, bids reflect the value of the set of consumers that search for certain keywords and not particular consumer histories. Bids are submitted per click and a payment from the advertiser to the intermediary occurs only if the consumer clicks on the advertisement. An important fact for the interpretation of our results is that consumers who opt out are never shown any advertisements. Thus, the intermediary generates no advertising revenues from these consumers.[36]

We separately investigate the changes in advertising revenue, prices, and quantity of advertisement. First, we look for the change in the number of clicks for advertisements following GDPR. Columns (1) - (2) of Table 3 show that there is a statistically significant decrease of 13.5% in the total number of clicks. The magnitude of this effect is in line with the drop in total cookies

[34] In the time period of interest, the new Microsoft Edge browser was the default on Windows computers and Internet Explorer is predominantly utilized on computers running the Windows OS. Microsoft Edge was the default on Windows since 2015, thus users of Internet Explorer are predominantly those on older computers. Internet Explorer users tend to be older than Chrome or Firefox users (`https://elie.net/blog/web/survey-internet-explorer-users-are-older-chrome-seduces-youth/`) and thus less inclined to adopt browser-based privacy practices (Zou et al., 2020).

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

[36] An implication of this is that although GDPR opt-out restricts the data observed by the intermediary and the website, we observe the advertising revenue for the intermediary generated from opt-in consumers.

and searches. We next look for changes in the number of clicks from distinct cookies to see if any changes were driven by some small set of consumers. Columns (3) - (4) show that this measure also decreases significantly. Figure 6 displays the time-varying specification for these outcome variable and shows that the effect on the number of clicks is relatively constant.

Columns (5) and (6) of Table 3 show the effects on revenue. The magnitude of the point estimates suggests an economically significant drop, though it is imprecise and 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 recover. Importantly, column (7) of Table 3 shows that the average bid of the advertisers *increases*. At roughly 12% this increase is economically sizable.[37] The time-varying coefficient in Figure 6 shows that the average bid does not change initially after the policy and then increases gradually. In summary, the immediate drop in clicks following GDPR leads to a sharp drop in revenue, but the gradual increase in the average bids leads to a recovery of some of the lost revenue for the intermediary and advertisers.

In light of these results one may wonder how the quantity of advertisements is affected. Using the same difference-in-differences specification with total number of advertisements as the dependent variable we find that the number of advertisements has dropped but that this change is not significant (see Table 10 and Figure 12 in Appendix E for the time varying treatment effect).

We now discuss the plausible mechanisms behind the increase in prices (bids). The first mechanism, which is consistent with the evidence that we establish above, is that remaining consumers are of higher average value to advertisers.[38] Our discussion with the intermediary indicates that advertisers' value is determined according to the *observed* conversion rate of their advertisements, which is the fraction of consumers that end up purchasing a good after clicking on an advertisement. Since the measurement of conversion rests on the ability to track consumers, it is plausible that the increased trackability of consumers following GDPR improved the measurement of conversion rates, thus contributing to an increase in value of consumers as perceived by the advertisers.[39] This points to the explanation that the increased ability to accurately measure conversion rates has led advertisers to gradually increase prices (bids) over time.

[37] Note that in order to preserve the privacy of our intermediary, the bid and revenue values are obfuscated by a common multiplier. However, the interpretation of percentage changes is preserved under this transformation.

[38] In contemporary work and in a different e-commerce setting, Goldberg, Johnson and Shriver (2021) reach a similar conclusion about the value of consumers post-GDPR.

[39] To illustrate, suppose that there are five consumers who click on an advertisement. Suppose one of them (from here on consumer A) makes us of cookie blockers 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 as opposed to 0.6—a correct rate including A. 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 regardless of consumer A's true behavior. More generally, dropping individuals similar to consumer A from the observed sample can only weakly increase the advertiser's perceived value.

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

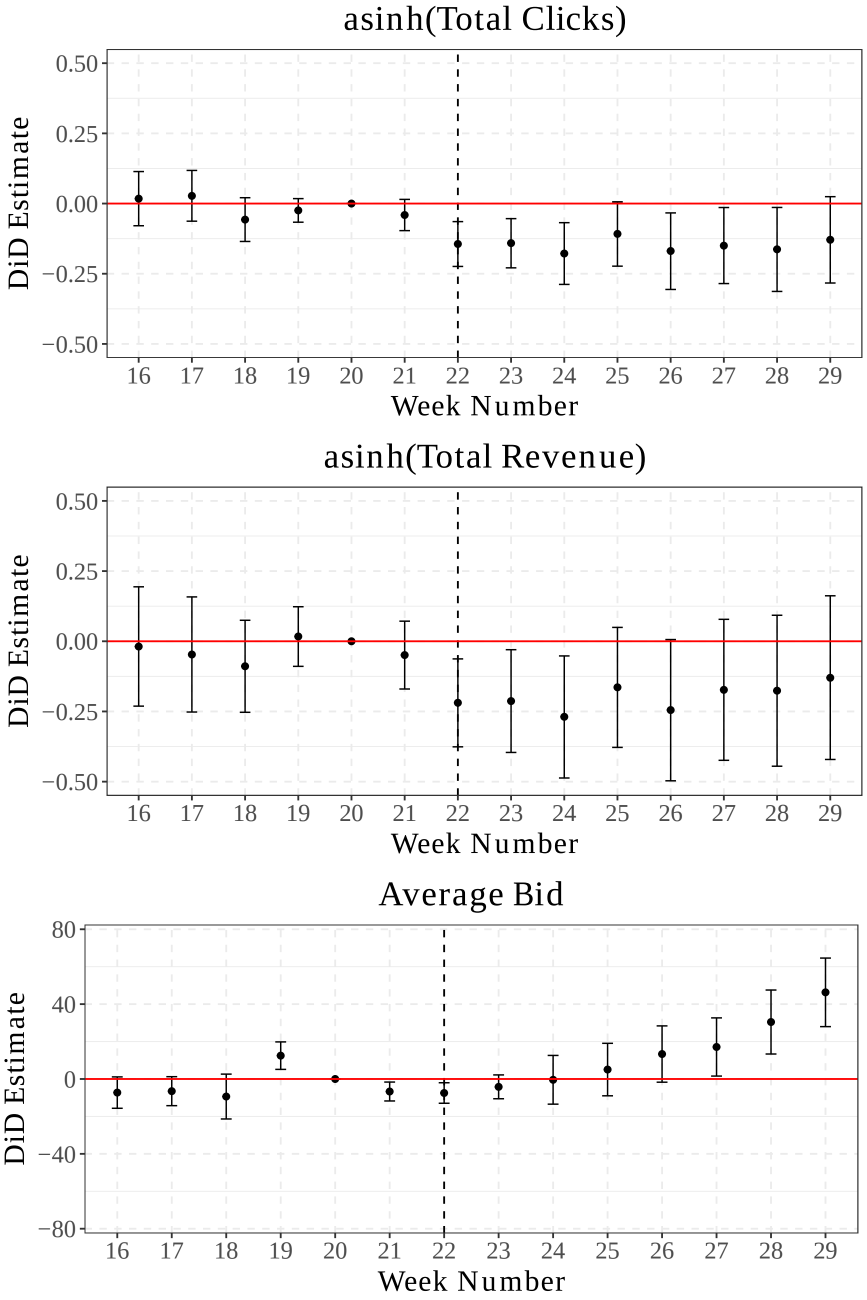


Table 3: Difference-in-Differences Estimates for Advertising Outcome Variables

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	asinh(Total	Total	asinh(Distinct	Distinct			Average
	Clicks)	Clicks	Clicks)	Clicks	asinh(Revenue)	Revenue	Bid
DiD Coefficient	-0.135**	-251.9*	-0.133**	-214.9*	-0.168	-32972.3	15.41***
	(-2.32)	(-1.91)	(-2.33)	(-1.84)	(-1.54)	(-0.75)	(2.90)
OS + Browser Controls	✓	✓	✓	✓	✓	✓	✓
Product Category Controls	✓	✓	✓	✓	✓	✓	✓
Website× Country FE	✓	✓	✓	✓	✓	✓	✓
Week FE	✓	✓	✓	✓	✓	✓	✓
Observations	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-browser-OS-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 bid by advertisers. We utilize the inverse hyperbolic sine transform instead of the logarithm as in previous sections as some of the outcome variables we consider in this section can take zero values. The inverse hyperbolic sine transform is given by $\bar{y} = arcsinh(y) = \ln(y + \sqrt{y^2 + 1})$ and results in a similar coefficient interpretation as taking logarithms (Bellemare and Wichman, 2019), but does not remove the zero valued observations from the data. We retain the zero values here so that there is a clearer comparison between the estimates before and after the transformation.

However, there might be two plausible alternative explanations. The first is that GDPR has decreased the "supply" of consumers to whom advertisements can be served. As we have demonstrated above, there is a significant reduction in the number of advertisements served because consumers opt out. This reduction in advertising targets might increase the value of the marginal remaining consumer. While this is certainly plausible, the pattern of price increases is not fully consistent with the supply shock explanation. Since this shock materializes right after the implementation date and advertising budgets are set daily, one would expect a sharp price increase. Instead, we observe a gradual price increase (Figure 6), which is more consistent with advertisers slowly adjusting to the increase in conversion rates.

Another plausible alternative explanation is that GDPR is a positive "demand shock" for the type of advertising offered by the intermediary. Advertisers in our setting submit bids based on the context in which advertising is shown (e.g. based on travel search details) instead of on individual consumer histories. The relative efficiency of such "contextual advertising" compared to behaviorally targeted display advertising, which is even more dependent on consumer tracking, may have increased as a result of GDPR.[40] However, our setting also contains a personalized element: the decision to place advertisements is personalized based on the predictions of the intermediary. It is therefore less plausible that the intermediary would be one of the clear-cut winners under such a shift in the market.

To sum up, it seems plausible that advertising prices, at least in part, increased because the average consumer is more trackable. This interpretation is in line with the evidence of previous sections and alternative explanations are less plausible based on our institutional knowledge and prevailing patterns in the data.

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. Beyond this particular context, such an investigation is also of broader interest. Sophisticated machine learning technologies that attempt to predict consumer purchase behavior are becoming increasingly common[41] and our results provide a case study on how their accuracy is affected by data privacy regulation.

Based on our analysis we expect there to be three predominant reasons why we might observe

[40] Personalization diminished: In the GDPR era, contextual targeting is making a comeback. https://digiday.com/media/personalization-diminished-gdpr-era-contextual-targeting-making-comeback/. Accessed on December 15th, 2020.

[41] See, for example, Retailers Use AI to Improve Online Recommendations for Shoppers, https://www.wsj.com/articles/retailers-use-ai-to-improve-online-recommendations-for-shoppers-11604330308, Accessed on March 31st, 2021.

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 consumer behavior and the length of consumer histories that were previously obfuscated by the use of alternative privacy means. We would expect the first effect to decrease prediction performance and the second and third to increase prediction 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 utilizing the 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 consumer i on website j on the kth query observed by the intermediary. We denote the classification made in real-time by the intermediary as $\hat{y}_{ijk}$. For every consumer 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 timestamp of when consumer i purchased on website j. This allows us to further construct the ground truth label, y_{ijk}^{TRUE}, which we use to evaluate the performance of the classifier.[42] We will denote the *class proportion* as the proportion of searches whose ground truth label is purchaser.

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

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

We observe the intermediary's predicted $\hat{p}_{ijk}$ and $\hat{y}_{ijk}$ for every search as well as the y_{ijk}^{TRUE} which we construct. The conversion of probability estimate, $\hat{p}_{ijk}$, to actual classification, $\hat{y}_{ijk}$, is based on whether the consumer's "score" $\hat{p}_{ijk}$ is above or below a chosen threshold $\hat{P}$. The threshold is chosen based on revenue considerations and other factors irrelevant to the quality

[42] The ground truth labels are constructed by setting $y_{ijk}^{TRUE} = 1$ if the purchase occurs within N_j days of the search and $y_{ijk}^{TRUE} = 0$ otherwise. While in practice the value of N_j is website-dependent, we do not observe this value for each website so we restrict focus to $N_j = 2$ across all websites. For the majority of websites in our sample ,the intermediary informed us that they set $N = 1$ or $N = 2$. Furthermore, from our preliminary analysis, the results do not qualitatively differ between $N = 1$ and $N = 2$.

of the predictions and, as a result, we focus on analyzing the prediction error associated with the probabilistic estimate $\hat{p}_{ijk}$ and not $\hat{y}_{ijk}$.

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).*[43]

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,

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

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.[44] The

[43] 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.

[44] We provide additional details on the construction of the AUC and its interpretation in Appendix F.1. For an

ROC curve in turn measures how well the classifier trades off Type I ("false positive") with Type II ("false negative") errors. The AUC 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 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 available only for a subset of websites.[45] 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. Since the pool of consumers has changed with GDPR our predictability regressions compare the larger set of consumers before GDPR with a smaller set of consumers after GDPR. Changes in predictability are therefore a function of both the quantity of data and the selection of consumers where consumers with longer histories remain in the data.

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 prediction probability (i.e. $\hat{p}_{ijk}$) in column (2) shows that little adjustment by the classifier of the firm to

extended discussion of ROC analysis, see Fawcett (2006).

[45] We drop observations that either have no purchase data or where the class proportion is degenerate. 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.

this change. Figure 14 in Appendix G 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.

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.[46] 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 the intermediary's ability to separate the two classes has increased. This observation is consistent with what we would expect from the aforementioned hypothesis of privacy means substitution.

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

	(1)	(2)	(3)	(4)	(5)	(6)
	Class	Average			Purchaser	Non-Purchaser
	Proportion	Predicted Probability	MSE	AUC	MSE	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-browser-OS-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.

[46] Appendix F.2 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.

Finally, Figure 15 in Appendix G 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 adjustment. Furthermore, 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 H 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 an increase in trackability will likely improve 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 trackability makes up for decreased data (resulting from opt-outs), as indicated by Appendix H, 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 magnitude 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 such regulation may put firms in these markets at a disadvantage relative to these larger firms. It would be important to study the extent and magnitude of these adverse effects. 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

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

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.

Batikas, Michail, Stefan Bechtold, Tobias Kretschmer, and Christian Peukert. 2020. “European Privacy Law and Global Markets for Data.”

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.*

Cameron, A Colin, and Pravin K Trivedi. 1990. “Regression-based tests for overdispersion in the Poisson model.” *Journal of econometrics*, 46(3): 347–364.

Cameron, A Colin, and Pravin K Trivedi. 2005. *Microeconometrics: methods and applications.* Cambridge university press.

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.

Coey, Dominic, and Michael Bailey. 2016. "People and cookies: Imperfect treatment assignment in online experiments." 1103–1111.

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.

Desmarais, Bruce A, and Jeffrey J Harden. 2013. "Testing for zero inflation in count models: Bias correction for the Vuong test." *The Stata Journal*, 13(4): 810–835.

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.

Einav, Liran, and Jonathan Levin. 2014. "Economics in the age of big data." *Science*, 346(6210): 1243089.

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.

Godinho de Matos, Miguel, and Idris Adjerid. 2019. "Consumer consent and firm targeting after GDPR: The case of a large telecom provider." *Management Science.*

Goldberg, Samuel, Garrett Johnson, and Scott Shriver. 2021. "Regulating Privacy Online: An Economic Evaluation of the GDPR." *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.

Jia, Jian, Ginger Zhe Jin, and Liad Wagman. 2020. "GDPR and the Localness of Venture Investment." *Available at SSRN 3436535.*

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

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.*

Johnson, Garrett, Scott Shriver, and Samuel Goldberg. 2020. "Privacy & market concentration: Intended & unintended consequences of the GDPR." *Available at SSRN 3477686.*

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.

Lambert, Diane. 1992. "Zero-inflated Poisson regression, with an application to defects in manufacturing." *Technometrics*, 34(1): 1–14.

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.*

Lin, Tesary, and Sanjog Misra. 2020. "The Identity Fragmentation Bias." *arXiv preprint arXiv:2008.12849.*

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.

Prince, Jeffrey, and Scott Wallsten. 2020. "How Much is Privacy Worth Around the World and Across Platforms?" *Available at SSRN.*

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.

Vuong, Quang H. 1989. "Likelihood ratio tests for model selection and non-nested hypotheses." *Econometrica: Journal of the Econometric Society*, 307–333.

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.

Zou, Yixin, Kevin Roundy, Acar Tamersoy, Saurabh Shintre, Johann Roturier, and Florian Schaub. 2020. "Examining the adoption and abandonment of security, privacy, and identity theft protection practices." 1–15.

Appendix

A Additional Consumer Response Figures

Figure 7: Week by Week Treatment Effect (Cookies and Recorded Searches)

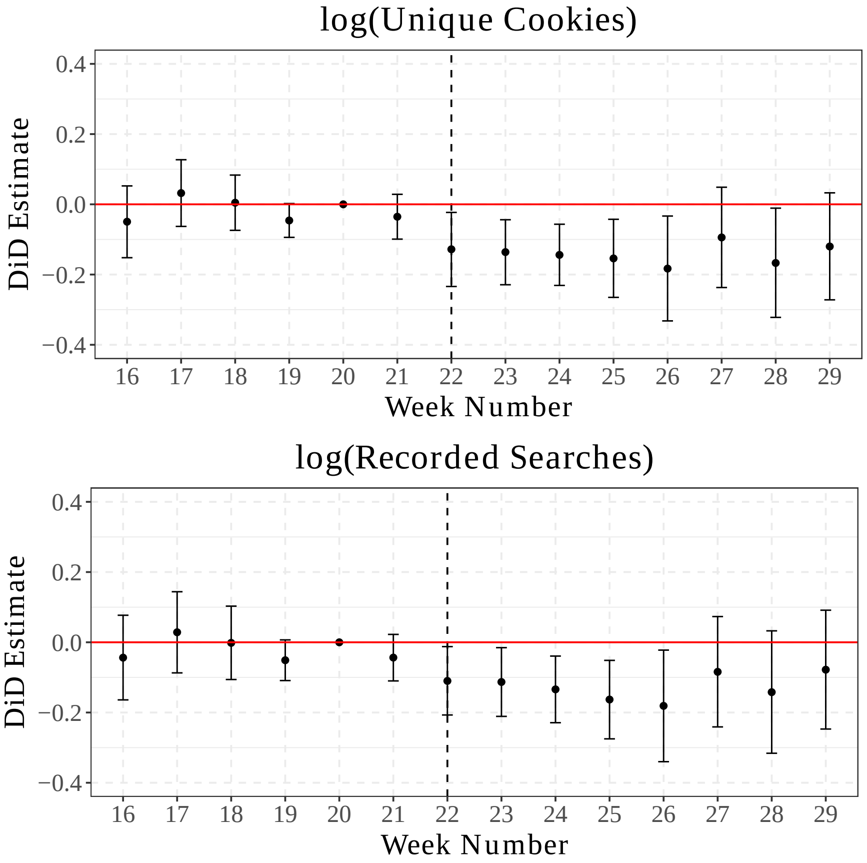


Table 5: Difference-in-Differences Estimates for Sales Activity

	(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 8: Week by Week Treatment Effect (Consumer Persistence)

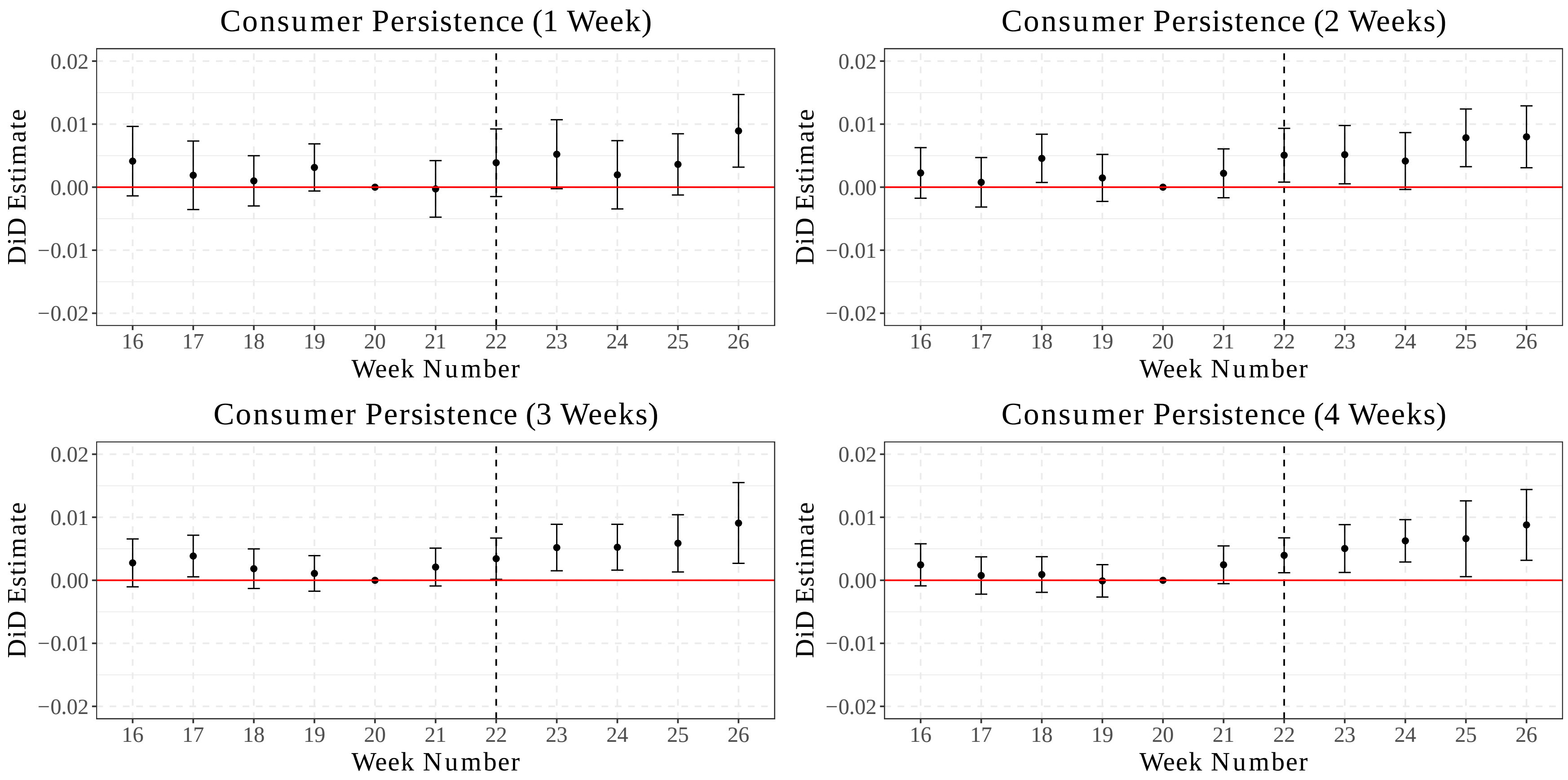
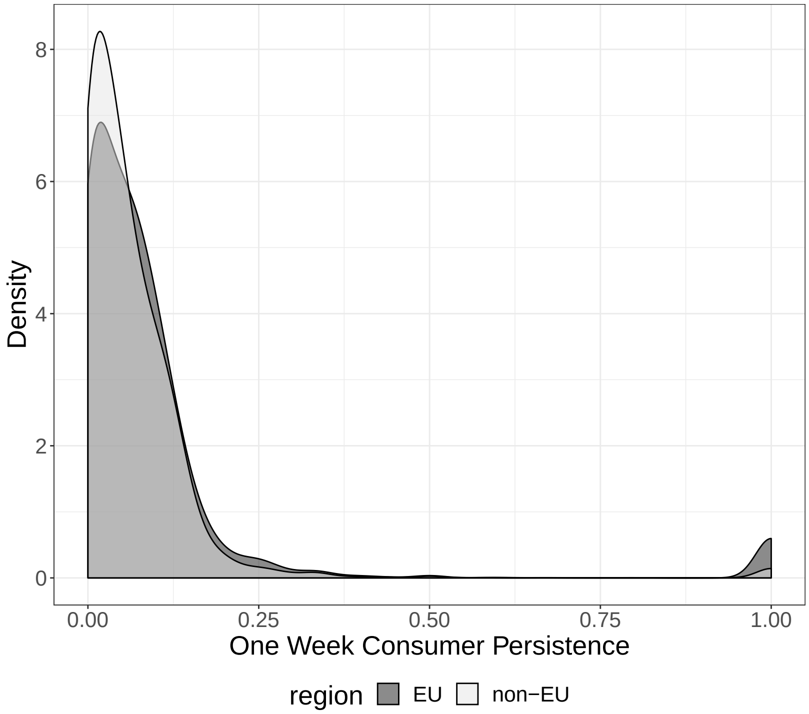


Table 6: Summary Statistics of Consumer Persistence

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

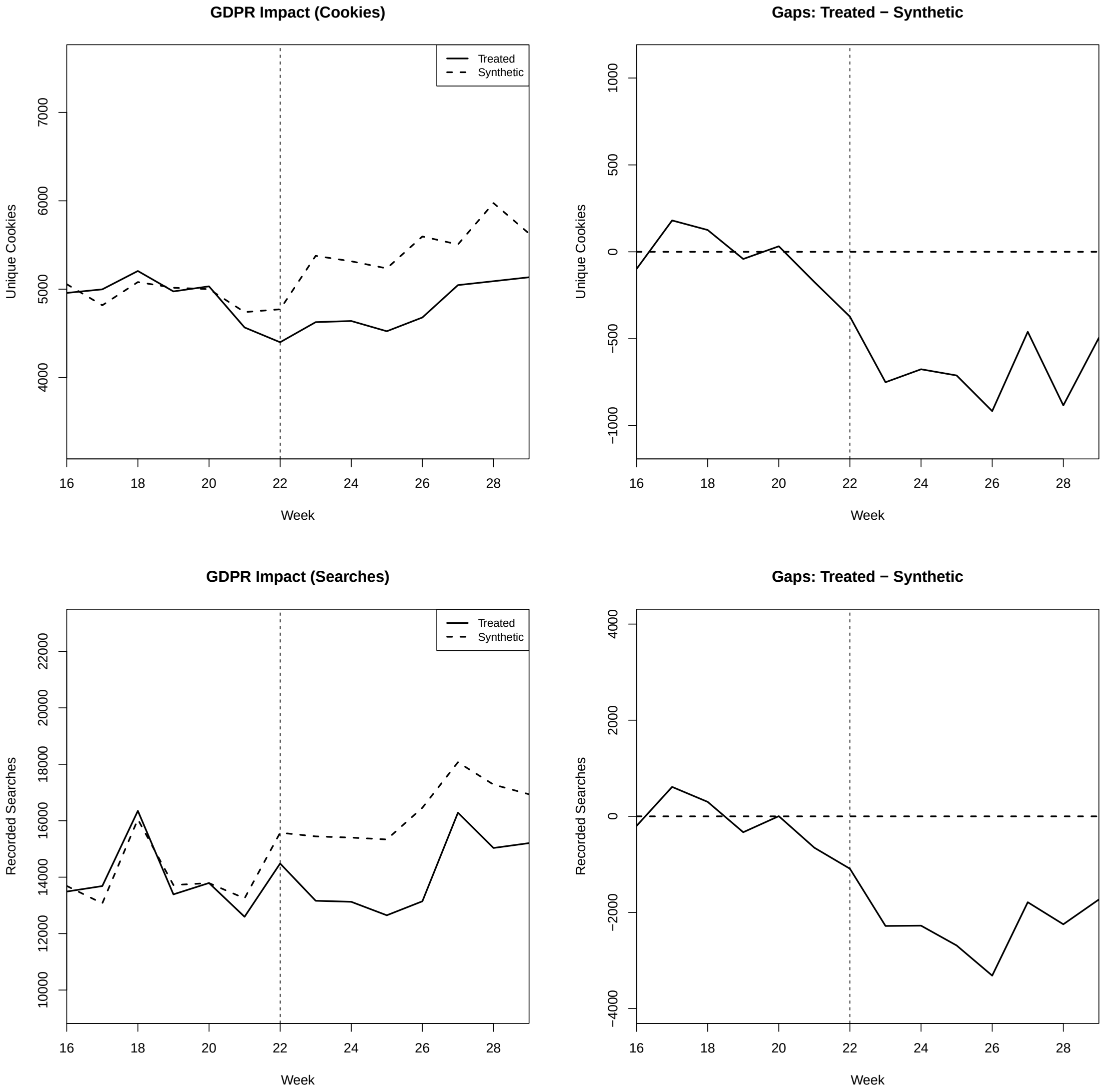
Notes: The summary statistics are computed on the sample period before GDPR and show the mean consumer persistence values across the EU and the non-EU for $k = 1, 2, 3, 4$.

Figure 9: Distribution of Consumer Persistence (1 Week)



B Robustness for Consumer Response Results

Figure 10: Synthetic Controls for Cookies and Recorded Searches



Notes: The plots in the leftmost column display the time series of the average treated unit and the constructed synthetic control for the number of unique cookies (top) and number of recorded searches (bottom). The plots in the rightmost column display the difference at every point in time between the averaged treated unit and the constructed synthetic control for the number of unique cookies (top) and number of recorded searches (bottom).

We provide additional evidence of robustness for our estimated effects on the usage of consent-based opt-out as a result of GDPR. subsection B.1 mimics the exercise in the main text, but utilizes a standard synthetic control based approach and recovers similar results as our difference-in-differences approach. subsection B.2 augments our analysis with Google Trends data and uses this to control for seasonality differences in travel patterns across the countries in our analysis.

B.1 Synthetic Controls

In order to provide additional validation for the difference-in-differences results in subsection 4.1, we supplement our primary analysis with a synthetic controls analysis, following Abadie, Diamond and Hainmueller (2010) and utilizing the corresponding R package, Synth. We aggregate the data to the same level as we do in the primary analysis.[47] We expand the set of control countries beyond the United States, Canada, and Russia to include Argentina, Brazil, Australia, Japan, and Mexico in order to allow for additional flexibility in the design of the synthetic control group.[48] Thus, the travel websites in these countries serve as the possible donor pool for the construction of the synthetic control. In order to apply the synthetic control method to our data we construct a single average treated unit for each outcome variable from the set of treated units. The set of predictor variables that we utilize in order to fit the weights assigned to each control unit are the two outcome variables that we consider—the total number of searches and the total number of unique cookies observed. We fit the weights to match the outcome variable between weeks 16 and 21. The results of applying this method are reported in Figure 10. Qualitatively, the results match what we find utilizing the difference-in-differences analysis with a stark drop at the onset of GDPR with a small recovery nearing the end of our sample period.

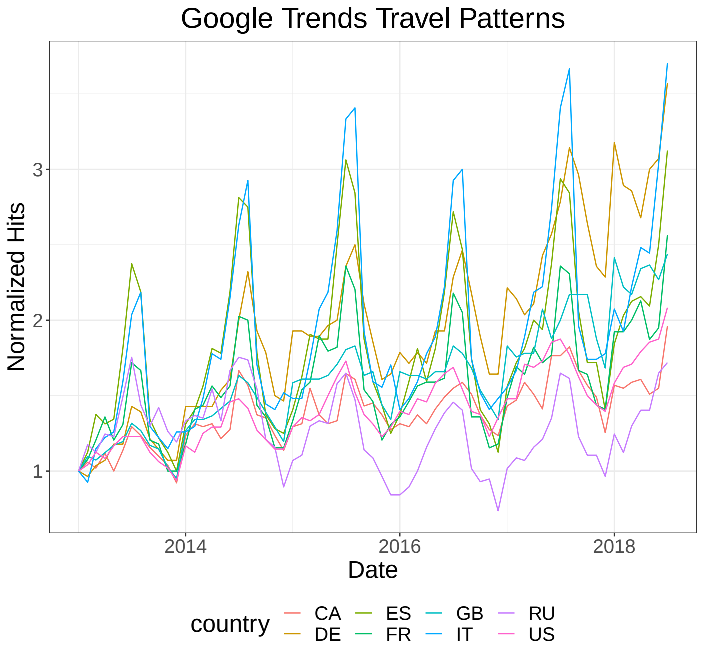
B.2 Controlling for Differences in Travel Patterns Across Countries

Since our paper tries to understand the impact of privacy regulations utilizing data from the online travel industry, a potential concern is that differential seasonality trends in travel across countries may influence the results. We selected the set of control countries in our analysis specifically to have similar seasonal travel patterns as the major EU countries impacted by GDPR in a short period around the GDPR implementation date. To further validate this we make use of data from Google Trends. The Google Trends data is useful since it provides an estimate of similar quantities observed in our data, but without the possibility that data can be removed as a result of GDPR. We first plot the relative trends over time of a common travel keyword and provide evidence that the travel trends are relatively similar in the period that we study. If anything, such trends seem to cause our estimates to understate the treatment effects of GDPR. We then make use of the historical data from Google Trends to better control for seasonal patterns and investigate the

[47] We also do this exercise aggregating at the website-country level and find qualitatively similar results. One might argue for this aggregation since the way we utilize the synthetic control method involves collapsing all treated units into a single average treated unit and it seems more natural to do so at the website-country level so that the synthetic control represents a synthetic European website. However, this makes the comparison of estimated treatment effects to the primary specification more difficult since the underlying units are different.

[48] Our results are nearly identical if we use the same set of control countries as we do in the baseline difference-in-differences specification, but use the larger set of countries due to the flexibility of the synthetic control method which makes this a special case of the reported exercise.

Figure 11: Historical Google Trends Travel Patterns



Notes: The graph is constructed by pulling Google Trends data for keyword "booking" for the time period ranging from 1/1/2013 - 7/31/2018. We pull the data for each country separately. We further normalize the score returned from Google Trends by dividing by the first observation for each country in order to ease cross-country comparisons.

impact of these controls on our estimates of the change in total recorded searches and unique cookies.

According to the Google Trends documentation, their data is constructed by a representative sample of searches done through Google Search. Instead of reporting the raw number of searches, Google Trends reports a normalized score that is constructed by dividing the number of searches for the keyword by the total searches of the selected geography and time range. The resulting number is scaled on a range of 0 to 100 based on the topic's proportion to all searches.[49]

Given this data construction, in order to compare the relative intensity of travel queries across countries we pull the data for each country and keyword individually. The first important detail is that Google Trends aggregates across specific strings and not terms, which means that when we do cross country comparisons we have to be careful about the precise keyword we utilize. In order to overcome this difficulty, we use the term of a common and popular OTA across all the countries in our analysis: booking. Figure 11 plots the results from Google Trends for the trends for this keyword from January 1st, 2013 until July 31st, 2018. Figure 11 shows that the keyword appears to pick up the seasonal trends we would expect across the different countries as well as that these appear to be similar across this set of countries, especially in the periods of

[49] `https://support.google.com/trends/answer/4365533` provides additional details.

our analysis.

Table 7: Difference-in-Differences Estimates With Google Trends controls

	(1) log(Unique Cookies)	(2) Unique Cookies	(3) log(Recorded Searches)	(4) Recorded Searches
DiD Coefficient	-0.129**	-1373.1*	-0.113**	-9555.9**
	(-2.52)	(-1.75)	(-1.98)	(-2.25)
Google Trends Seasonality Controls	✓	✓	✓	✓
OS + Browser Controls	✓	✓	✓	✓
Product Category Controls	✓	✓	✓	✓
Week FE	✓	✓	✓	✓
Website × Country FE	✓	✓	✓	✓
Observations	63840	63840	63840	63840

t statistics in parentheses

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

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-browser-OS-product type-week level between weeks 16 and 26, including both weeks 16 and 26 (April 13th - June 29th). These regressions are identical to those presented in the main text, but with the addition of Google Trends data in order to control for potential differences in seasonal travel patterns across the countries in our analysis.

We now consider the same specification as in (1), but make use of the Google Trends data to additionally construct controls for seasonal travel trends. We run the following regression in order to construct these controls, using the daily Google Trends data from 2013-2018:[50]

$$google_{ct} = \chi\Big[week \times country\Big] + \epsilon_{ct} \tag{5}$$

where as in the main specification, t denotes week and c denotes country. We then take $\hat{\chi}$ and

[50] For this analysis we aggregate the daily Google Trends normalized scores to a weekly level. We define a week in an identical manner as in the primary analysis, from Friday-to-Friday, and take the average normalized score over the week in order to construct this data.

add into our primary specification:

$$y_{tcjobp} = \alpha_t + \hat{\chi}_{tc} + \delta_{jc} + \gamma_o + \zeta_b + \omega_p + \beta(EU_j \times after) + \epsilon_{tcjobp} \tag{6}$$

where the notation is identical to that utilized in the main text and $\hat{\chi}_{tc}$ denotes the coefficient on $week \times country$ that comes from from running (5). The regression results are reported in Table 7 and are qualitatively consistent with the results from our main specification.

C 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:

$$y_{tcjobp} = \alpha_t + \delta_{jc} + \gamma_o + \zeta_b + \omega_p + \beta(EU_j \times after \times browser) + \epsilon_{tcjobp} \tag{7}$$

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 are 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.[51] The lack of JavaScript extensions on Internet Explorer makes cookie blockers substantially more difficult to implement on the browser

[51] Safari also is the default browser on OS X and so one would expect users to potentially be less technically sophisticated than those that make use of non-default browsers.

and thus we would expect less “single searchers” and less usage of browser-based privacy means due to the relative lack of automated means of doing so.[52] In sum, in order to be consistent with the privacy means substitution hypothesis we would expect a smaller increase in persistence on Internet Explorer and Safari relative to the other browsers.

Table 9 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, consistent with our hypothesis. The estimated treatment effect is lower in Safari relative to Chrome, but the difference is not statistically significant. Both of these observations are consistent with the privacy means 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:

$$y_{tcjobp} = \alpha_t + \delta_{jc} + \gamma_o + \zeta_b + \omega_p + \beta(EU_j \times after \times OS) + \epsilon_{tcjobp} \tag{8}$$

We are mainly interested in differences in the treatment effects between mobile and desktop consumers. The reason is that 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 a smaller difference.

Table 8 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 $k = 3, 4$. 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 be suggestive of the privacy means substitution hypothesis, but does not provide conclusive evidence.

[52] See, for instance, https://help.getadblock.com/support/solutions/articles/6000055833-is-adblock-available-for-internet-explorer-.

Table 8: 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-browser-OS-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 9: 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-browser-OS-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.

D Additional Evidence for “Single Searcher” Inflation

In this section we more formally investigate the “single searcher” observation from subsection 4.2. Our main objective is to determine statistically whether there is an excess amount of single searchers, which is consistent with the privacy means substitution hypothesis. The test results in this section provide evidence for the following two claims. First, consumers were making use of such privacy means prior to the introduction of GDPR. Second, the fraction of consumers doing so has decreased after GDPR is introduced.

D.1 Setup and Hypotheses

We first describe a simple model that motivates our empirical exercise. Suppose that there are two types of consumers – obfuscators (o) and non-obfuscators (n) – and that each type of consumer generates an observed search history length of k. Each consumer has a history of length $k \geq 1$. For the obfuscation type o, we hypothesize that their observed history length is $k = 1$ with probability one. For the non-obfuscation consumer type n, the probability of observing search history length $k \geq 1$, conditional on observing a consumer, is denoted by $Q(k; \theta(x))$, where θ contains the relevant parameters of probability distribution Q and x denotes a set of observable consumer characteristics. In our setting the lowest count is one, which is why we subtract one from each observation to map it into a standard count model. We denote the fraction of consumers that are obfuscators conditional on observable characteristics x as

$$\pi(x) := \Pr\{\text{obfuscator} \mid x\}$$

This setup maps into the following observed share of visits S_k, where k denotes history length:

$$S_1 = \pi(x) + (1 - \pi(x)) \cdot Q(1; \theta(x)) \tag{9}$$

$$S_k = (1 - \pi(x)) \cdot Q(k; \theta(x)), \quad \forall k \geq 1 \tag{10}$$

We note that given this set-up π and θ are identified and that we can separately estimate π and θ for the pre-GDPR and post-GDPR period, giving us estimates for $\hat{\pi}^{PRE}, \hat{\pi}^{POST}, \hat{\theta}^{PRE}, \hat{\theta}^{POST}$. Given this setup, our informal hypotheses can be stated as the following null hypotheses:

1. $H_0 : \hat{\pi}^{PRE} = 0$, $H_a : \hat{\pi}^{PRE} \neq 0$
2. $H_0 : \hat{\pi}^{POST} = \hat{\pi}^{PRE}$, $H_a : \hat{\pi}^{POST} \neq \hat{\pi}^{PRE}$

D.2 Data and Estimation

For this exercise we restrict attention to the same large hotel website shown in Figure 5 (which exhibited the noticeable change in "single searchers" at the onset of GDPR). We measure how many searches are associated with each identifier observed before and after GDPR is introduced. In total, we observe more than three million unique identifiers.

We allow the parameters of the model to depend on both the web browser and the operating system. Thus, we allow both the arrival rates and the fraction of obfuscators to vary across these dimensions. Next, we parameterize $\pi(x)$ as follows:

$$\pi(x) = \Big[\exp(x'\gamma)\Big] / \Big[1 + \exp(x'\gamma)\Big],$$

where γ is a parameter to be estimated. We consider two possible distributional assumptions for Q: a Poisson distribution and a negative binomial distribution, where the latter allows for additional dispersion. For the Poisson distribution we allow the arrival rate $\lambda(x)$ to vary across observables and we do similarly for the negative binomial parameters $\mu(x)$, $\alpha(x)$.[53]

Our setup maps almost directly to standard zero-inflation Poisson models (e.g. Lambert (1992). We follow Lambert (1992); Cameron and Trivedi (2005) and estimate the parameters of the model via maximum likelihood estimation. The model with a positive share of obfuscators is tested against either a standard Poisson regression or a negative binomial regression. We then conduct a Vuong test (Vuong, 1989) to evaluate whether a model with type o consumers leads to a better fit to the observed data (Desmarais and Harden, 2013). In order to test our second hypothesis of interest, we do a t-test comparing the vectors of $\hat{\pi}^{PRE}$ and $\hat{\pi}^{POST}$.

D.3 Results

We first consider the specification where we assume that Q follows a Poisson distribution. The results of the Vuong test strongly conclude that there is evidence for the existence of type o consumers in both periods with a z-statistic of -244.85 in the pre-GDPR period and -246.28 in the post-GDPR period.

We then compare the resulting $\hat{\pi}$ in the pre-GDPR and post-GDPR periods, denoted by $\hat{\pi}^{PRE}$ and $\hat{\pi}^{POST}$, respectively. We run a t-test with the null hypothesis that $\hat{\pi}^{POST} = \hat{\pi}^{PRE}$. We are able to reject the null with $p < 2.2e - 16$. The difference is also economically significant as we note that $\overline{\hat{\pi}^{PRE}} = 0.478$ and $\overline{\hat{\pi}^{POST}} = 0.354$, suggesting a significant drop of obfuscators after GDPR.

[53] See section 20.4.1 of Cameron and Trivedi (2005) for full details of the parameterization for Poisson and Negative Binomial regressions that we utilize.

One concern with the parameterization of Q as Poisson is that it does not account for overdispersion or underdispersion. We can directly test for overdispersion. Let Y_i denote the observed history length for consumer i and $\hat{\lambda}_i$ the implied variance of the poisson distribution. One can then test the null that $\alpha = 0$ for $\texttt{VAR}(Y_i) = \hat{\lambda}_i + \alpha \cdot \hat{\lambda}_i$ against the alternative that α is larger than zero (Cameron and Trivedi, 1990). We reject the null ($p < 2.2e - 16$) in both the pre and post period. Thus, we conclude that the data is overdispersed and consider the common remedy that imposes that Q follows a negative binomial distribution, instead of a Poisson distribution (Cameron and Trivedi, 2005).

Under the the assumption of a negative binomial, the Vuong test still concludes that there is evidence for the presence of type o consumers ($z = -6.97$) in the pre-GDPR period. However, we no longer reject the model without excess single searchers in the post-GDPR period ($z = -1.81$, AIC-corrected: $z = -0.84$, BIC-corrected: $z = 4.94$). Furthermore, we are able again to reject the null hypothesis that $\hat{\pi}^{POST} = \hat{\pi}^{PRE}$ with $p < 2.2e - 16$.

In sum, we document statistical evidence for excess "single-searchers" in the pre-GDPR period under both distributional assumptions. Once we take into account the overdispersion relative to a Poisson count model, we do not find evidence for excess single searchers in the post-GDPR period.

E Additional Advertisement and Auction Figures

Table 10: Difference-in-Differences Estimates for Advertisements Delivered

	(1) Total Advertisements Delivered	(2) asinh(Total Advertisements Delivered)
DiD Coefficient	-2627.2 (-1.61)	-0.145 (-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-browser-OS-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 12: Week by Week Treatment Effect (Total Advertisements Delivered)

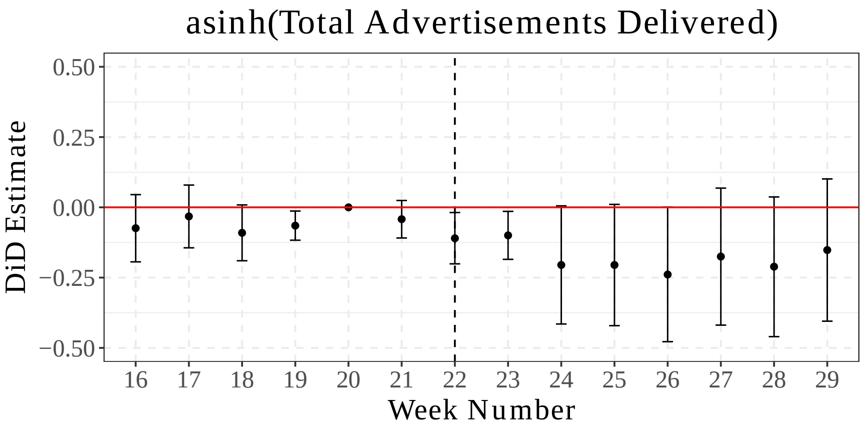


Table 11: Summary Statistics, Bids

Treatment Group	**Average Bid**
non-EU	394.053
EU	126.947

Notes: The table reports the mean of the average bid across observations in the pre-GDPR time period for the EU and non-EU respectively.

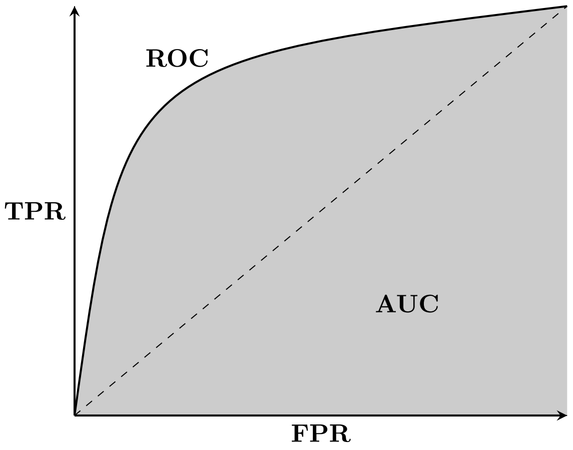
F Prediction Evaluation Measures

F.1 AUC Primer

In this section we provide additional details on how to calculate the AUC measure and its interpretation. 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 result in a false positive rate—a rate at which a non-purchaser is misclassified into a purchaser:

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

Figure 13: 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.

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:

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

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 $\hat{P}$.[54] 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 13 depicts a typical ROC curve.

F.2 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

[54] 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 $(FPR, TPR) = (1, 1)$.

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.[55] For notation, p_k denotes the kth estimated 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:

$$MSE_j = \underbrace{\frac{1}{\sum\limits_{i \in \mathcal{I}_j} |\mathcal{K}_{ij}|} \sum_{k=1}^{K} n_k (p_k - \bar{o}_k)^2}_{\text{calibration error}} + \underbrace{\frac{1}{\sum\limits_{i \in \mathcal{I}_j} |\mathcal{K}_{ij}|} \sum_{k=1}^{K} n_k \bar{o}_k (1 - \bar{o}_k)}_{\text{refinement error}} \tag{11}$$

We run the same specification utilizing each component of the decomposition of the MSE as the outcome variable. These results are reported in Table 12. 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.).

[55] 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.

Table 12: Difference-in-Differences Estimates for Relevance and Calibration

	(1) Calibration	(2) Refinement
DiD Coefficient	0.00735*** (2.84)	0.00576** (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-browser-OS-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 Additional Prediction Figures

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

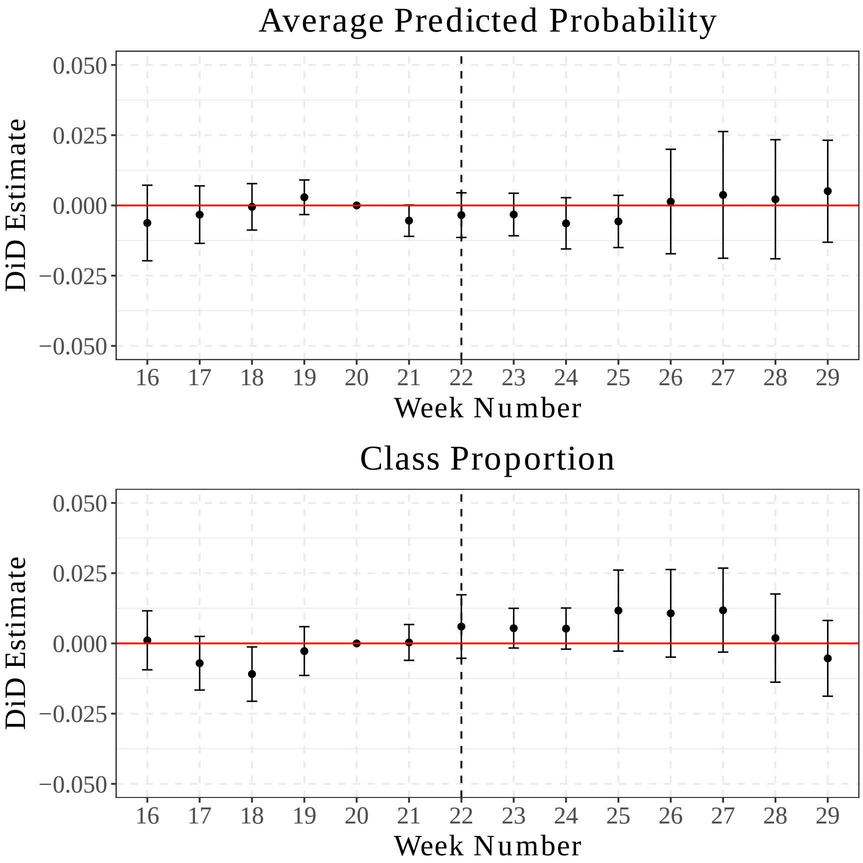
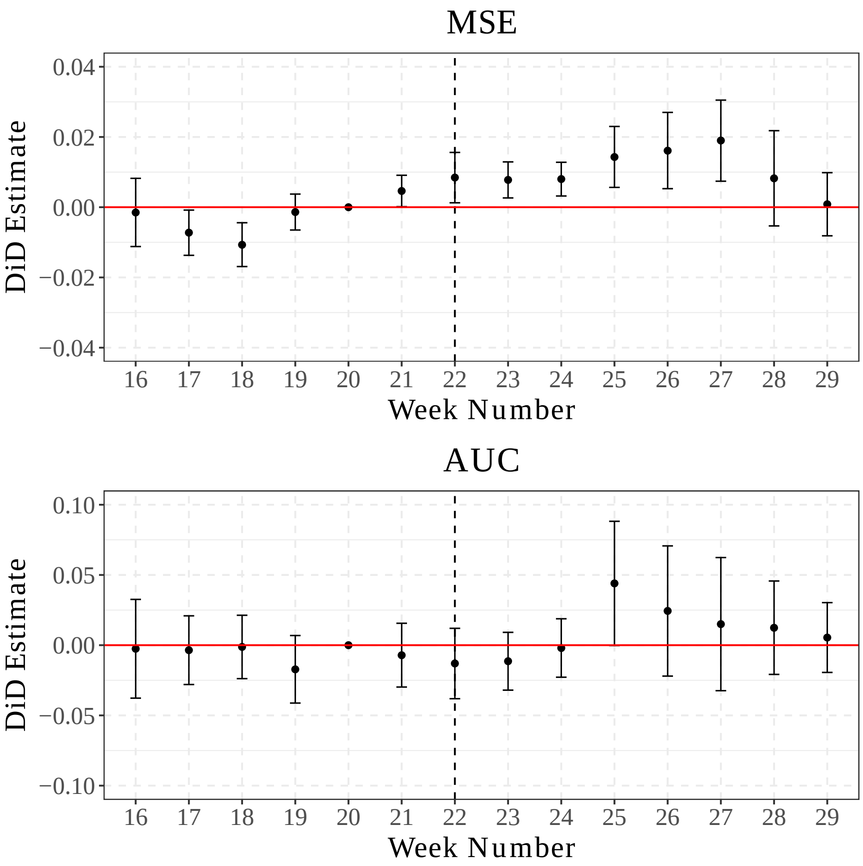


Figure 15: Week by Week Treatment Effect (MSE and AUC)



H 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:

$$pred_{tcjp} = \beta \cdot \log(Recorded\_Searches) + \alpha_t + \delta_{jc} + \omega_p + \epsilon_{tcjobp} \tag{12}$$

$$pred_{tcjp} = \beta \cdot Consumer\_Persistence + \alpha_t + \delta_{jc} + \omega_p + \epsilon_{tcjobp} \tag{13}$$

Table 13 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.[56] This corresponds to the regression results in Columns (1) and (3) of Table 13. 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 with a 10.7% decrease in recorded searches only leading to a 0.0007 decrease in AUC.[57]

Table 15 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 15, 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

[56] 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.

[57] 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 14 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.

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.

Table 13: 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 14: 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 15: 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.