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

Guy Aridor† Yeon-Koo Che‡ Tobias Salz§

September 28, 2022

Abstract

Utilizing a novel dataset from an online travel intermediary, we study the effects of the EU's General Data Protection Regulation (GDPR). The opt-in requirement of GDPR resulted in a 12.5% drop in the intermediary-observed consumers, but the remaining consumers are trackable for a longer period of time. Our findings imply that privacy conscious consumers exert privacy externalities on opt-in consumers, making them more predictable. Consistent with this finding, 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 article was previously circulated as *The Economic Consequences of Data Privacy Regulation: Empirical Evidence from the GDPR*. We are grateful to the editor Katja Seim and three anonymous referees for excellent feedback that greatly improved the article. We would further like to thank Daron Acemoglu, Francesco Decarolis, Glenn Ellison, Sara Ellison, Avi Goldfarb, and seminar and conference participants at Columbia University, ASSA Meetings, the MaCCI/EPoS Virtual IO Seminar, Columbia Data Science Day, FTC PrivacyCon, Yale Law Big Tech & Antitrust Conference, the ACM Conference on Economics and Computation, BIRS Workshop on Statistical Methods for Computational Advertising, Toulouse Economics of Platforms Seminar, and 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.

†Northwestern University, Kellogg School of Management. Email: guy.aridor@kellogg.northwestern.edu.
‡Columbia University. Email: yeonkooche@gmail.com.
§Massachusetts Institute of Technology. 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. However, at the same time, there has been an increasing number of high profile data breaches, such as the Cambridge Analytica scandal, and a growing feeling of despondency amongst consumers who lack control over this process.[1] Beyond the immediate economic harm resulting from such data breaches consumers might also value privacy for its own sake.[2] 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 many other countries and US states.[3] 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 article 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

[1] We Hate Data Collection. That Doesn't Mean We Can Stop it. New York Times Privacy Survey. Retrieved on January 3rd, 2020.

[2] For the different motivations for privacy see Acquisti et al. (2016). As noted by Lin (2022), consumer privacy preferences contain both an instrumental and non-instrumental component.

[3] Those states and countries include, for instance, Brazil, California, Chile, Colorado, India, Maine, Nevada, New Zealand, Utah, Vermont, and Virginia. For more information on the specifics of the various laws and how they relate to GDPR: Privacy Laws Around The Globe. Retrieved on March 10th, 2020.

consumer behavior and target their services and advertising accordingly.

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, it has been argued that consumers may not be willing to act on their privacy concerns even at little cost or inconvenience.[4] We do not yet have clear empirical answers to this question, and consumers' GDPR opt-out decisions 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. However, utilizing these privacy means does not eliminate their footprints altogether but rather simply generates "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. Moreover, consumers may selectively consent based on how loyal and trusting they are to the websites, which could increase the fraction of more persistently identifiable consumers. 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 et al., 2019; Acemoglu et al., 2019; Bergemann et al., 2022) 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?

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

These questions are particularly important for the competitive landscape of the digital economy. Although 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.[5] 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. How such a (relative to the internet giants) smaller third-party firm copes with GDPR could provide a valuable clue on how data regulation may influence the competitive playing field of the digital economy.

To answer these questions, we use 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 for the host website. 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 host 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 *across the host online travel agencies* as the intermediary and link them just as the intermediary can. If a consumer does not consent to data sharing 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

[5]This concern has been raised in the recent literature (Johnson et al., 2020; Peukert et al., 2022) and popular press (Wall Street Journal - GDPR Has Been a Boon for Google and Facebook. Retrieved on June 2nd, 2020) which show that GDPR led to an increase in market concentration of web trackers, favoring those from Google and Facebook.

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 affected 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. It is important to clarify that our analysis characterizes the overall impact of the policy. There are documented cases of non-compliance (DPC, 2020b) and GDPR may have changed the salience of privacy around its implementation. Our estimates should therefore be viewed in the context of this partial compliance and increased salience.

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 two plausible mechanisms for the increased trackability. The first 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. Although 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. The second is that consumers may selectively consent; namely, less frequent users of a website are less trusting of its privacy protection. Due

to a combination of these mechanisms, the consumers that remain in the data after the implementation of GDPR are those who 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 to the advertisers, compared with the pre-GDPR set of consumers. Of the two possible mechanisms for increased trackability, this is more consistent with consumers substituting away from obfuscation to opt-out, because the opt-out of obfuscators with short search spells 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 the ability of the classifier to separate between purchasers and non-purchasers did not significantly worsen after GDPR. If anything, additional analysis suggests that the intermediaries ability to separate between purchasers and non-purchasers should improve in the long run.

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. The combination of selective consent and switching away from less efficient browser-based means of privacy protection to 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. 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; although 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 also be 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.[6]

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 et al. (2021). We see these two studies as complementary in terms of the data scenario and findings. Our study utilizes data at a more dis-aggregate level but is confined to one industry whereas Goldberg et al. (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 articles have studied the impact of the GDPR in other domains (Jia et al., 2018, 2020; Zhuo et al., 2021; Utz et al., 2019; Degeling et al., 2018). Peukert et al. (2022); Johnson et al. (2020) show that GDPR increased market concentration amongst web technology services. Goldfarb and Tucker (2011); Johnson et al. (2020) study the effectiveness of previous data privacy regulations on online advertising. Godinho de Matos and Adjerid (2021) conduct an experiment with

[6]This could occur through personalized pricing. See, for instance, Dubé and Misra (2019) and Buchholz et al. (2022), for welfare quantifications of such pricing policies.

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 articles 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.[7] Several recent theoretical studies argue how such externalities can lead to the underpricing of data, and results in socially excessive data collection (Fairfield and Engel, 2015; Choi et al., 2019; Acemoglu et al., 2019; Bergemann et al., 2022; 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 article identifies a novel form of informational externalities. Whereas 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 externalities that run in the opposite direction. Namely, we show that the decision by a privacy-concerned consumer 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 article that identifies privacy externalities empirically.[8]

Preferences for Privacy: The broader literature on the economics of privacy, recently surveyed in Acquisti et al. (2016), has studied the privacy preferences of individuals. One prevalent research strand is understanding the privacy paradox, which is the apparent disparity between

[7] There is also an emerging, broadly related, literature that studies implications of a more data-driven economy (Goldfarb and Tucker, 2012a; Einav and Levin, 2014; Chiou and Tucker, 2017; Kehoe et al., 2018; Aridor et al., 2020; Bajari et al., 2019)

[8] 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, 2022).

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 et al., 2005; Norberg et al., 2007; Athey et al., 2017). Acquisti et al. (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 (2012b) show that consumer's privacy concerns have been increasing over time. Lin (2022) shows via a lab experiment that consumer privacy preferences can be broken down into instrumental and non-instrumental components. Our study contributes to this literature by analyzing consumer privacy choices made in a consequential setting, instead of only looking at stated preferences. We find that a significant fraction of consumers utilize the privacy means provided by GDPR, giving suggestive evidence that consumers do value their privacy in consequential settings and not only say that they do.

The article 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 and Conceptual Framework

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. Furthermore, we will comment on how different mechanisms for opt out may affect the advertising market and the ability of the intermediary to predict consumer purchase behavior.

European Data Privacy Regulation

The GDPR was adopted by the European Parliament in April 2016. Companies were expected to comply with the new regulations by May 25th, 2018.[9] The law required wide-ranging changes in how firms collect, store, and process consumer data. The fines for non-compliance with the rules are large - the maximum is set at €20 million, or 4% of total global annual sales for the preceding financial year - giving strong incentives for firms to comply with the regulation, which led them to spend millions of dollars to do so.[10]

We focus on the consumer-facing aspect of the regulation, namely the conditions under which a firm can collect consumer data. Regulators have clarified that the most appropriate legal basis for data collection, especially for the type of data collected by the intermediary, is individual consent (DPC, 2020a). From Recital 32 of 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:

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

Panel (a) of Figure 1 shows an example of a commonly used post-GDPR cookie policy from Quantcast and panel (b) of Figure 1 shows a cookie policy of a firm in the United States. The

[9] GDPR was intended to replace the Data Protection Directive and complements the existing Privacy and Electronic Communications Directive. Relative to the latter, GDPR strengthened the territorial scope to include data generated by EU consumers and strengthened the degree of firm transparency and stipulations on user consent.

[10] Pulse Survey - GDPR budgets top $10 million for 40% of surveyed companies. Retrieved on December 15th, 2019.

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. By contrast, 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 allows them to control the website's cookies. The intermediary that we partner with is a third-party affiliate and, for the typical consent notices, is never explicitly mentioned by name. Thus, it is likely the identity of the OTA, rather than the intermediary itself, is more salient for consumer's opt-in choice in our setting.

Potential Effects on Consumers and Firms

GDPR enabled consumers to opt out of data-sharing. Although consumers may opt-out for various reasons, we focus on two plausible ones here. We then discuss their implications for the patterns of observed consumer behavior and their consequences for firms and advertisers. Throughout, these two mechanisms will be used to interpret our empirical findings.

Selective Consent. The first is *selective consent*, where consumers only consent to data processing by websites that they trust and thus frequently use. According to this mechanism, infrequent users of a website are more likely to opt out of data sharing than frequent users. If this is the primary mechanism for opting out, then one direct consequence is that privacy regulation may favor bigger or more reputed firms. There is a connection of this mechanism to the theoretical predictions in Campbell et al. (2015), who argue that consent-based data collection practices would allow larger firms to collect more data than smaller firms because they offer a wider scope of services. As a result, consumers may visit the websites more and share more data with it. In the long run, this would imply that consent for data collection may serve as a barrier to entry.

This is illustrated in Figure 2. The figure shows the data generated by three consumers. "Full Visibility" shows a hypothetical scenario where each of the three consumers is fully identifiable. The right panel under GDPR shows that the infrequent consumer – consumer 3 – opts out and is no longer observed by the intermediary. From the intermediary's perspective, the remaining

consumers will be seen to be more persistently identifiable.

The figure also illustrates how selective consent can impact the intermediary's ability to predict whether a consumer will purchase from the given website. Consider the intermediary's prediction problem for consumer 2 in the full visibility baseline. At $t = 3$ its "training data" consists of a mixture of an "infrequent" consumer – consumer 3 – and a "frequent" consumer – consumer 1 – whose data in their first observed search is identical. Thus, the ability of the intermediary to predict whether consumer 2 will ever purchase a flight is more difficult because her search history pools with both types, despite being a "purchaser" type. However, under the GDPR regime, the infrequent consumers opt out of the data, which makes the prediction problem of the intermediary for consumer 2 easier. The key aspect that drives the information externality of consumer 3 is dynamic in nature – if early in the search process the "infrequent" types pool together with the "frequent" types, then GDPR opt out makes it easier for the intermediary to separate the two.

Finally, selective consent would have consequences for the advertising market in our setting. Recall that the intermediary targets the consumers that are deemed less loyal to the host website and thus are more likely to respond to competing OTA's ads. If the opt-out consumers are the infrequent users of the host websites, as suggested by the theory, then the remaining consumers are likely to be those loyal to the host websites and thus unlikely to respond to the competing OTAs' ads. This means that the average value of consumers is likely to decrease after GDPR. Moreover, consumers according to this mechanism only share data with their preferred OTA making it harder to win them over with advertisements for comparison shopping, which is the main kind of advertisement of the intermediary. Hence, we would expect advertisers to spend less on advertisements for comparison shopping and prices to decrease.

Substitution of the means for protecting privacy. The GDPR opt-out presents consumers with a powerful way of protecting their privacy. For privacy-concerned consumers, this means that they can substitute from other less effective privacy means to simple opt-out. Prior to GDPR,

such consumers were likely to employ a variety of methods to protect themselves against consumer tracking by websites. We will denote this type of substitution *privacy means substitution*.

One prominent method that websites, including our intermediary, use to track consumers is through web cookies.[11] 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. Privacy-conscious consumers can use 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,[12] or advertising/cookie blockers,[13] all of which we will call *obfuscation* throughout. One important detail to note is how advertising/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 although 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 but the data that is generated by them on the website is still sent and stored. Hence, a consumer's data will be attributed to different identifiers.[14]

The stipulations of GDPR, properly implemented and utilized by consumers, arguably provide a stronger protection than the aforementioned means because they block all non-essential information from being sent to the third-party website. It is important to note that GDPR does not prevent "essential" information from being sent to a website. For instance, the ability to store

[11]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.

[12]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 session.

[13]There also exist industry opt-out services, such as the Ad Choices program but these have low take up (Johnson et al., 2020). Based on survey evidence, the most utilized manual privacy means is cookie deletion (Boerman et al., 2018). Furthermore, there is wide adoption of advertising and cookie blockers (see e.g. Statista – Ad Block Users WorldWide. Retrieved on April 4th, 2022 – which reports 732.32 million users of ad blockers worldwide in 2018).

[14]One possible instrumental reason why consumers in this setting would engage in privacy-preserving behavior is that there is a belief, erroneous or not, that cookies are used to implement personalized pricing (see e.g. Business Insider - Clear Cookies When Searching for Flights. Retrieved on April 4th, 2022).

consumer session cookies that allows 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. In our context, this means that by simply opting out consumers can keep their data from being sent to the intermediary because it provides a non-essential, third-party service. In principle, consumers may still prefer other means to protect their privacy for fear that firms otherwise disadvantage them. However, the GDPR does not allow firms to discriminate against consumers who do not consent to data sharing, making opt-out more attractive all else equal.[15] Hence, it is plausible that privacy-conscious consumers would substitute from obfuscation to opt-out.

Such substitution will have a distinct effect on the data observed by the intermediary. Before GDPR, reliance on obfuscation by a consumer will mean that her data would still be sent to the intermediary but with many "bogus" identifiers associated with each visit she makes. After GDPR, if such a consumer simply opts out, as postulated by the current mechanism, no data from 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.

This is illustrated in Figure 3. 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 4 is privacy-conscious. Before GDPR, consumer 4 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.

[15] In that respect it is more stringent than, for instance, the CCPA that allows firms to set explicit incentives for data sharing. See Cal. Civ. Code § 1798.125 for the CCPA's stipulation and Baker Law - GDPR vs. CCPA Chart (Retrieved on April 6th, 2022) for a comparison of the GDPR and the CCPA.

However, the third panel shows that, when GDPR opt-out is available, this consumer opts out and her 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, because it cannot tell who obfuscates and who does not. Under GDPR, on the other hand, consumer 4's data is not observed at all. Although this leads to a loss in the amount of data, consumer 4 no longer confounds the prediction for consumer 1 and 2's behavior.[16]

This last point has an implication for the advertisements mediated by the intermediary. As noted earlier, the intermediary targets advertisements to consumers with short spells of cookie histories at the host website, which importantly includes the obfuscators. Prior to GDPR, the obfuscators, with new cookies artificially generated for every visit including one that lead to a purchase, are not easily linked across ad clicks and purchases. In other words, prior to GDPR, a significant presence of the obfuscators makes it difficult for advertisers to "attribute" purchases

[16]Importantly, the pre-GDPR intermediary cannot simply replicate the same dataset as post-GDPR because 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.

to ad clicks. Hence, their presence undermines the value of ad clicks as perceived by advertisers.[17] Suppose, as postulated by the mechanism, a significant number of obfuscators opt out after GDPR. Then, the number of consumers with short spells of cookie histories will decrease, so the number of consumers that the intermediary targets for advertising will fall. However, the remaining consumers targeted by the intermediary—namely the opt-in consumers with short spells of cookie histories—are more likely to be linked across their web visits, so their purchases are more easily attributed to ad clicks. This will increase the average value of consumers (targeted by the intermediary) to the advertisers and increase prices.

3 Data and Empirical Strategy

We obtained access to a new and comprehensive dataset from an anonymous intermediary that records the entirety of consumer search queries and purchases across most major online travel agencies (OTAs) in the United States and Europe as well as most prominent travel meta-search engines. 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.

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

[17] To illustrate, suppose that there are five consumers who click on an advertisement. Suppose one of them (from here on consumer A) makes use 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.

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.

Overall, we observe, in aggregated form, data on the searches, advertisements, and purchases that occur on each of the online travel agencies and travel meta-search engines that the intermediary contracts with. Our analysis utilizes an aggregation of this dataset by week, operating system, web browser, website identifier, product type, and country. We drop from this aggregation observations which the intermediary has identified as bots. The data is 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.[18]

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 able to comply with the regulation by asking consumers to consent to data sharing, although this was not necessary for non-EU consumers. Even though many online travel companies transact with consumers in several countries around the world this specification works well in our setting because it is common for online travel websites to have separate, country-specific versions of their websites and only the websites intended for EU

[18] We enforce a balanced panel by dropping any observation that has zero logged searches in any week during our sample period in order to ensure that our estimates are not biased from entry / exit. This is usually due to varying contractual relations between the intermediary and the websites and so is orthogonal to our variables of interest.

countries are made GDPR compliant.

Our analysis captures the overall effect of the policy and not the specific effect of consent implementation. 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 because 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:

$$y_{t,c,j,o,b,p} = \alpha_t + \delta_{j,c} + \gamma_o + \zeta_b + \omega_p + \beta \cdot (EU_j \times after_t) + \epsilon_{t,c,j,o,b,p} \tag{1}$$

EU_j denotes a website subject to the regulation, $after_t$ denotes whether the current week is after the GDPR compliance date (i.e. week 22 or later), α_t denotes time fixed effects, $\delta_{j,c}$ 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.[19]

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:

$$y_{t,c,j,o,b,p} = \alpha_t + \delta_{j,c} + \gamma_o + \zeta_b + \omega_p + \sum_{k=\underline{T}}^{\bar{T}} \beta_k \cdot EU_j + \epsilon_{t,c,j,o,b,p} \tag{2}$$

[19]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).

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 because there are some firms that began to implement GDPR near the end of week 21 and so week 20 is the last week where there should be no direct impact from GDPR as a result of website implementation.[20]

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.[21] Thus, any effects that we observe with our empirical specification are a combination of the explicit consequences 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. As a result, our estimates can be viewed as measuring the overall impact of the policy.

[20] Some of our measures require at least two full weeks of data and so we drop the last full week and the incomplete week at the end of July. Our analysis ends at the end of July because we were not able to obtain any data after that.

[21] Many of the consent dialogs for GDPR are dynamically generated and thus not always captured by tools such as the Wayback Machine. However, we were able to verify that several websites did and others did not implement GDPR consent guidelines around the time of the policy, though for many we are uncertain about their implementation. We do not drop the non-compliant websites because we are broadly interested in the overall impact of the policy.

4 Consumer Response to GDPR

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

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.[22]

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

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 $U_{jt}^{OPT-OUT}$.[23,24]

Figure 4 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

[22]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.

[23]To our knowledge there is no change in the data the websites send to the intermediary as a result of GDPR because the intermediary and the data are crucial for the websites' advertising revenue. Furthermore, if a website decided to stop using the intermediary altogether then, as noted previously, they would not be part of our sample.

[24]Another possible confounding factor is varying sales activity of the intermediary. To test this hypothesis, we repeat our analysis with total advertising units and advertising pages as the outcome variables. The results in Table 6 show that there was no significant change in either of those two variables.

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 by Figure 4.

Another measure of consumer response is the total number of searches that are recorded 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 8 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 Online Appendix A.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 multi-national 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 4. 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, 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, because 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 Online Appendix A.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:

$$\text{persistence}_{k,t} = \frac{|C_{j,t} \cap C_{j,t+k}|}{|C_{j,t}|}$$

In Figure 5 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 $k \in \{1, 2, 3, 4\}$.[25,26]

Table 2 shows the results of this regression, which indicate that there is a statistically signifi-

[25]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.

[26]Note that the units on the regression and Figure 5 are not the same. Figure 5 displays the persistence measure in terms of percent deviations from week 20 whereas the coefficients in Table 2 are changes in levels.

cant and meaningful increase in consumer persistence and that this effect gets more pronounced as k increases.[27] 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 10 shows that although 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.[28] The treatment effect remains roughly the same as k grows, even though Table 5 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.

Recall from section 2 that there are two plausible mechanisms for driving this increase in persistence: privacy means substitution and selective consent. Although it is hard to fully disentangle the two explanations, we are able to provide suggestive evidence that both play a role. For that we focus on a large website that has faithfully implemented a consent policy. Although both mechanisms imply that the drop in relative probability mass should be concentrated at the lower end of the support, in our context one signature of browser-based privacy protection is a large mass of "single search" consumers. This is because advertising/cookie blockers continually regenerate the identifier after every search, as discussed in section 2. Indeed, Figure 6 shows that the fraction of single searchers significantly dropped after the implementation of GDPR and the relative probability mass of searchers with more than one search increased. Although a disproportionate drop in single searchers is suggestive of a larger role for privacy means substitution, it is still plausible that part of this could be driven by selective consent.

To test how privacy means substitution and selective consent contribute to the observed changes, we set up a simple model (for details see Appendix B). The model generates the distribution of search histories as a mixture of obfuscators, who always generate search histories of length one, and non-obfuscators, whose history is captured by a parametric count distribution.

[27] 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.

[28] Figure 9 shows the overall distribution of consumer persistence for the EU vs. non-EU and note that there are some outliers. Our results are qualitatively robust to winsorizing and dropping these observations as well. They are also robust to the addition of seasonal travel controls using the same procedure as in Online Appendix A.2.

Under this formalization, the "excess" number of single searchers identifies the fraction of obfuscators and the residual count distribution the true underlying search histories. We can use a Vuong test to determine whether obfuscators are needed to fit the observed distribution (Vuong, 1989). We use the test to determine the presence of obfuscators and whether the fraction changes with the introduction of GDPR. In addition, we can test for whether true search histories are significantly longer after GDPR comes into effect, which is indicative of selective consent.

We estimate the model under two different distributional assumptions for the count distribution of the true number of searches. First, we assume a conditional Poisson distribution and, second, we assume that it is negative binomial, to allow for more dispersion. In short, we find evidence for both of the mechanisms as the test suggests the presence of obfuscators in the pre-GDPR period but not in the post-GDPR period, although also showing that the average length of the (true) underlying search history increased slightly.

We can additionally test for the presence of selective consent by investigating whether large and small websites exhibit heterogeneous treatment effects.[29] Selective consent implies that persistence should increase more for large websites (e.g. consistent with Campbell et al. (2015)). The results of this regression are reported in Table 7, which points to a larger, but not statistically significant, increase in persistence for larger websites. We interpret this as weak evidence for the presence of selective consent.

Finally, we investigate 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 suggestive evidence that it is more pronounced on desktop compared to mobile operating systems. These differences are plausible under privacy means substitution, in light of the alleged lack of technical sophistication of Internet Explorer users and the technical difficulty of utilizing browser-based privacy means on mobile devices and the Internet Explorer.[30] However,

[29] We define a "large" and "small" website as a website above (resp. below) the median size. Recall from section 2 that the consent decision is elicited with no information about the intermediary. As a result, the most relevant degree of heterogeneity is the size of the first party website.

[30] The Microsoft Edge browser was the default on Windows computers since 2015 and thus users of Internet Explorer are predominantly those on older Windows computers. Furthermore, Internet Explorer users tend to be older than Chrome or Firefox users (Internet Explorer Users are Older – Retrieved on May 27th, 2022) and thus less

the results between other browsers and operating systems are less clear cut which suggests that selective consent is also driving the results. The results and a full discussion are deferred to Online Appendix B.

Overall, these results provide robust evidence for privacy means substitution driving the increase in persistence, although also some degree of selective consent.

5 GDPR and Online Advertising

The advertisements in our setting are sold by the intermediary on the host OTA website via real-time auctions that are held when a consumer makes a search query.[31] Advertisers, who are typically competing OTAs or airlines, 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 same travel itinerary. Thus, bids are shaped by the average value of attracting a consumer to the advertisers' website for the same itinerary.[32] 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.[33]

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

inclined to adopt browser-based privacy practices (Zou et al., 2020).

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

[32] The fact that the advertisers are bidding to direct the consumer to their website for the same search differs from the standard sponsored search auction in which the same product may be sold under many keyword searches and categories, which makes conversion and attribution difficult. In particular, it means that unlike standard keyword search advertising, it is clear whether an ad converted into purchase so long as a third-party cookie allows the advertiser to link the attributed purchase to the advertisement placed on the host website. Thus, the only possible changes to the ability to do attribution in this setting would be through changes in the ability to track the average consumer. Importantly, this means that advertisers do not bid on particular consumer histories, so any observed changes come from a change in the overall composition of consumers.

[33] 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.

GDPR. Columns (1) and (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 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) and (4) show that this measure also decreases significantly. Figure 7 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 7 shows that revenue initially falls sharply after the implementation of GDPR and then begins to slightly recover. Importantly, column (7) of Table 3 shows that the average bid of the advertisers *increases*. At roughly 12% this increase is economically sizable.[34] The time-varying coefficient in Figure 7 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 8 and Figure 11 in Appendix A for the time varying treatment effect).

We now discuss the plausible mechanisms behind the increase in prices (winning 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.[35] Our discussion with the intermediary indicates that advertisers' value is determined according to the *observed* conversion rate of their

[34]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. The average obfuscated bid value in the EU pre-GDPR was 126.9.

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

advertisements, which is the fraction of consumers that end up purchasing a good after clicking on an advertisement. Consistent with our explanation in section 2, it is plausible that as a result of privacy means substitution 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. It is further important to recall from section 2 that the observed increase in prices is inconsistent with the predictions of selective consent.

There are several pieces of evidence supporting privacy means substitution. First, before GDPR 24.9% of advertisements on flight searches are targeting consumers whose cookie history is of length one, which indicates that this group is quantitatively important for advertisers. Second, we find that the fraction of consumers searching for flights who receive advertisements has increased, which is consistent with the explanation that the average remaining consumer has become more valuable to advertisers (see Table 9).[36] One possible objection to this interpretation may be that the two types of consumers systematically differ in their search and purchase patterns, resulting in differences in value to advertisers based on the types of flights that they buy. However, we have no indication that consumers with single cookie histories search for flights of different value. For instance, we find that their average days to departure when searching and fraction of searches that are for domestic flights are extremely similar to consumers with longer histories (see Table 10 and Table 11). Thus, it is likely that the set of remaining consumers are equally valuable and the price increase is due to a higher chance of attribution.

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 advertisement to the remaining consumers. Although this is certainly plausible, the pattern of price increases is not fully consistent with the supply shock explanation. As this shock materializes right after the implementation date and advertising budgets are set daily, one would expect a sharp

[36]Recall that due to the nature of targeting by our intermediary selective consent would imply that remaining consumers are less valuable to advertisers and thus less likely to receive advertisements.

price increase. Instead, we observe a gradual price increase (Figure 7), which is more consistent with advertisers slowly adjusting to the equally slow 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.[37] 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. We also provide two pieces of direct evidence that the change in prices is not predominantly driven by changes in demand. First, we find that the number of bidders in the advertising auctions of the intermediary has not increased and in fact slightly decreased (Column (5) in Table 13). Although this rules out a positive demand shock at the extensive margin (i.e. from advertisers reallocating across different ad media), it could still be that specific advertisers purchase more advertisements than usual, perhaps because of a big campaign. In order to explore whether this was the case we compute the share of winning bids for each advertiser and construct common measures of buyer concentration – HHI and concentration ratios. Although we find a statistically significant increase in concentration, this effect is economically small (Columns (1)-(4) in Table 13). For instance, the increase in market concentration of the largest five advertisers per website (CR-5) is roughly 3% (see Table 12).

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. However, we must caution that the evidence that we provide is not conclusive about the mechanism and that it is known that inference in such niche advertising

[37] Digiday - Personalization diminished: In the GDPR era, contextual targeting is making a comeback. Retrieved on December 15th, 2020.

markets may be driven by large advertisers. Nonetheless, our results in this section suggest the importance of understanding the interplay between advertising attribution and privacy regulation.

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 and our results provide a case study on how their accuracy is affected by data privacy regulation.[38]

Based on our analysis we expect there to be three predominant reasons why we might observe a change in the ability to predict. First, GDPR has significantly reduced the overall amount of data. Second, remaining consumers have longer histories and are more trackable. Third, in line with our illustrations in Figure 2 and Figure 3, 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 the current site based on the history that the intermediary observes about this consumer on the current website. 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

[38] See, for example, Wall Street Journal - Retailers Use AI to Improve Online Recommendations for Shoppers. Retrieved on March 31st, 2021.

within some time window. Formally, each query is classified into

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

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 $\hat{y}_{i,j,k}$. For every consumer i we observe a series of searches on website j, $X_{i,j,1}, X_{i,j,2}, ..., X_{i,j,n}$ 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, $y_{i,j,k}^{TRUE}$, which we use to evaluate the performance of the classifier.[39] 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:

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

We observe the intermediary's predicted $\hat{p}_{i,j,k}$ and $\hat{y}_{i,j,k}$ for every search as well as the $y_{i,j,k}^{TRUE}$ which we construct. The conversion of probability estimate, $\hat{p}_{i,j,k}$, to actual classification, $\hat{y}_{i,j,k}$, is based on whether the consumer's "score", $\hat{p}_{i,j,k}$, is above or below a chosen threshold $\hat{P}$. The threshold is chosen based on revenue considerations and other factors irrelevant to the quality of the predictions and, as a result, we focus on analyzing the prediction error associated with the probabilistic estimate $\hat{p}_{i,j,k}$ and not $\hat{y}_{i,j,k}$.

[39] The ground truth labels are constructed by setting $y_{i,j,k}^{TRUE} = 1$ if the purchase occurs within N_j days of the search and $y_{i,j,k}^{TRUE} = 0$ otherwise. The intermediary typically sets $N_j = 1$ or $N_j = 2$. As we do not observe this value, our baseline specification uses $N_j = 2$ for all j, but our results do not qualitatively differ when $N_j = 1$.

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)*.

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

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

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 $\hat{p}_{i,j,k}$ 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.[40] The ROC curve in turn measures how well the classifier trades off Type I ("false positive") with Type

[40] We provide additional details on the construction of the AUC and its interpretation in Online Appendix C.1.

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 because they provide two qualitatively different measures of prediction performance.

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.[41] 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. As 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

[41] We drop observations that either have no purchase data or where the class proportion is degenerate as well as two websites that had a reporting error for purchase data during our sample period. Furthermore, we drop any $(browser, OS, product, website, country)$ tuple that, on average, has fewer than 50 consumers a week.

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.[42] Meanwhile, the insignificant coefficient for average prediction probability (i.e. $\hat{p}_{i,j,k}$) in column (2) shows little adjustment by the classifier of the firm to this change. Figure 12 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.[43] 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 increase in consumer persistence.

Finally, Figure 13 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.

[42] To interpret this recall that this is the set of purchasers on website j and not the set of purchasers more generally. Thus, this provides some additional evidence for a weak effect of selective consent in the remaining set of consumers because it implies that the average remaining consumer post-GDPR is more likely to purchase on website j.

[43] Online Appendix C.2 decomposes the change of MSE to account 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.

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 Online Appendix D 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 article 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 effectively required firms to ask consumers for explicit consent to store and process their data. Despite only partial compliance with this aspect of the regulation, we find a meaningful change in the data collected by firms.

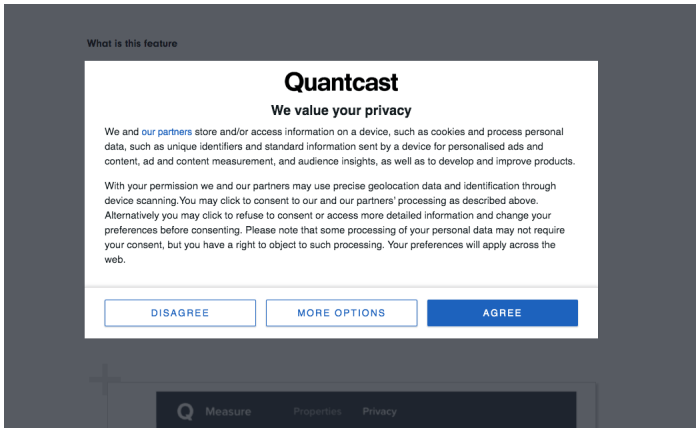
Our results paint a novel and interesting picture of how a consumer's privacy decision— particularly the means by which she protects her privacy—may affect the rest of the economy, including other consumers, and the firms and advertisers relying on consumer data. The strong and effective means of privacy protection made available by laws such as GDPR and the more 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. 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.

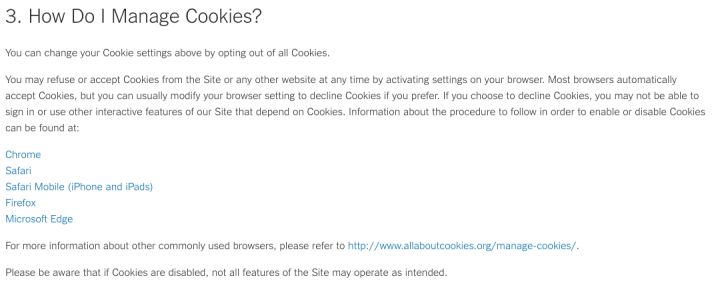
Regarding firms, we find two effects that partially offset each other. On the one hand, we find that advertising prices increase, which can be explained by the increased trackability of consumers. However, our analysis also shows that smaller advertisers such as ours, who are dependent on third party access, are able to collect less data and conduct less business due to consumer opt-out. This puts them at a disadvantage relative to large companies who rely on their first party ecosystem to collect data. Thus, our article has broader implications beyond the online travel industry and keyword-based advertising markets. Firms in this industry, as others in the digital economy, increasingly compete with the large technology firms such as Google whose reach spans across many different online markets and for whom consumers have little choice but to accept data processing. Thus, although our results highlight that increased consent requirements may not be wholly negative, 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 smaller firms at a disadvantage relative to the internet giants. 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.

Figure 1: Example Consent Notifications

(a) Post-GDPR consent dialog

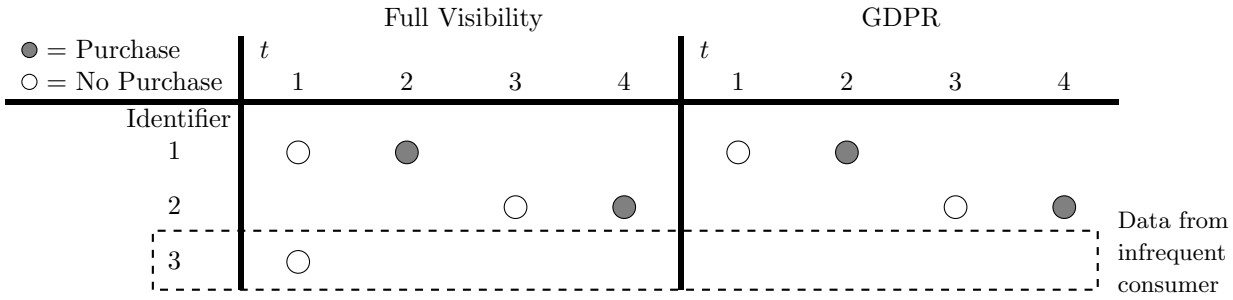


(b) Standard opt-out on US websites



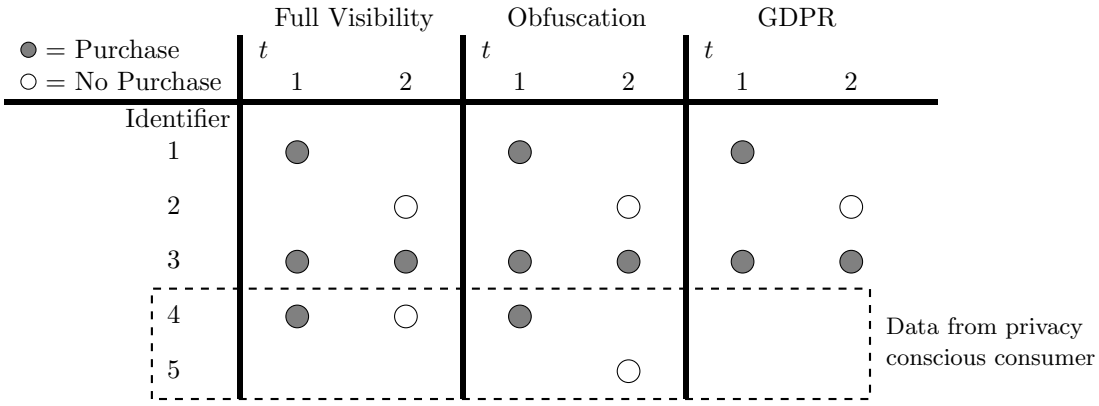
Notes: The top panel shows a standard GDPR opt in consent dialog provided by Quantcast. The dialog is explicit about the data that the website collects and requires the consumer to opt into all non-essential data collection. The bottom panel shows an "opt out" dialog for a website in the US that is not 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.

Figure 2: Illustration of Effects of Selective Consent on Data Observed



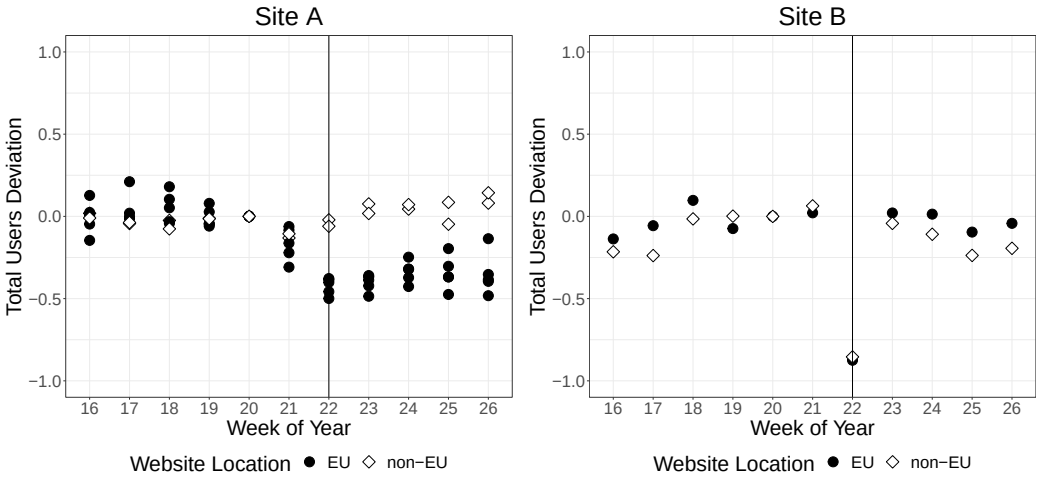
Notes: The leftmost column displays the identifier observed by the intermediary. The left panel represents the scenario where the behavior of each consumer is fully observable. The right panel shows how, under GDPR, the data of the infrequent consumer 3 is not directly sent to the intermediary.

Figure 3: Illustration of Effects of Obfuscation on Data Observed



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

Figure 4: 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 verified implemented the consent guidelines. 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 verified did not implement the consent guidelines. The black dots represent the values from the United Kingdom and the white dots represent the values from the United States.

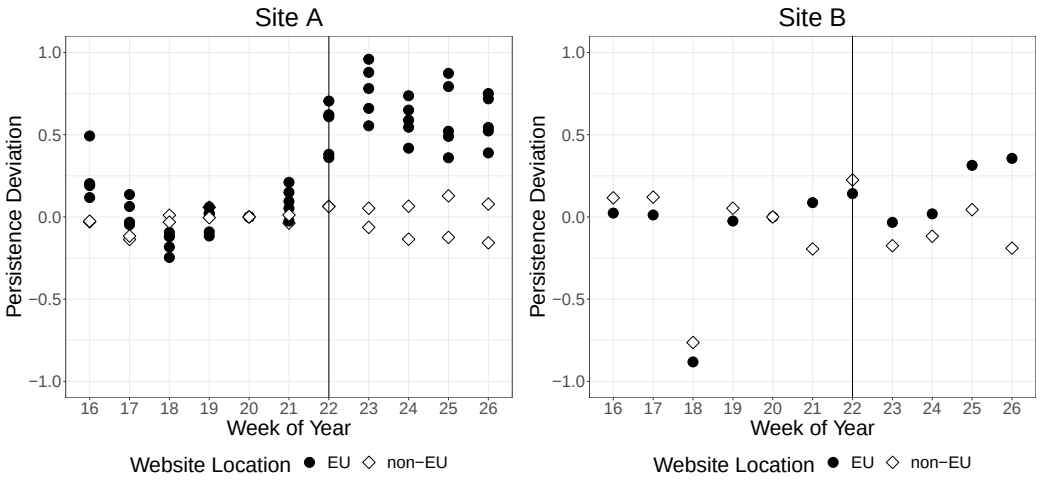
Table 1: Difference-in-Differences Estimates for Cookies and Searches

	(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.

Figure 5: 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 verified implemented the consent guidelines. 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 verified did not implement the consent guidelines. The black dots represent the values from the United Kingdom and the white dots represent the values from the United States.

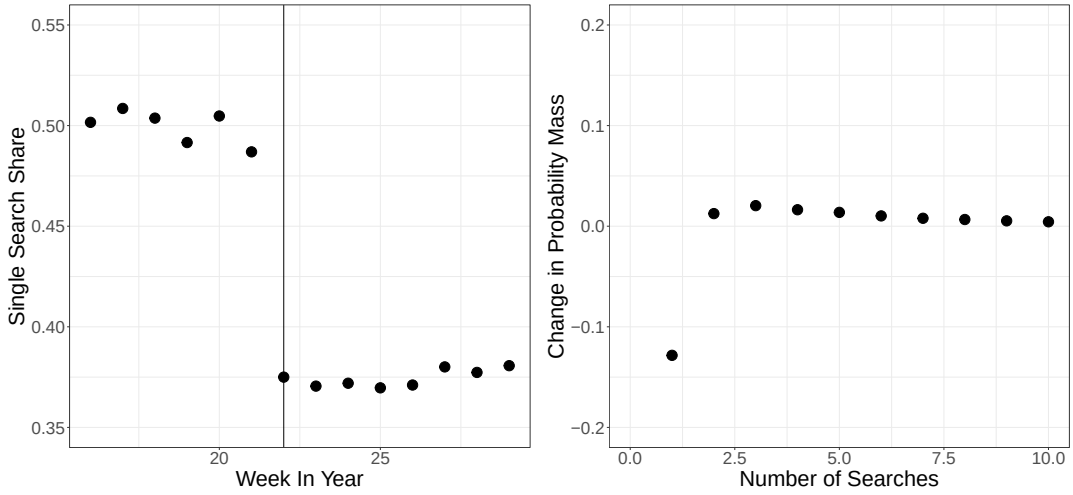
Table 2: Difference-in-Differences Estimates for Consumer Persistence

	(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.

Figure 6: 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. 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.

Figure 7: Week by Week Treatment Effect for Total Clicks, Revenue, and Average Bid

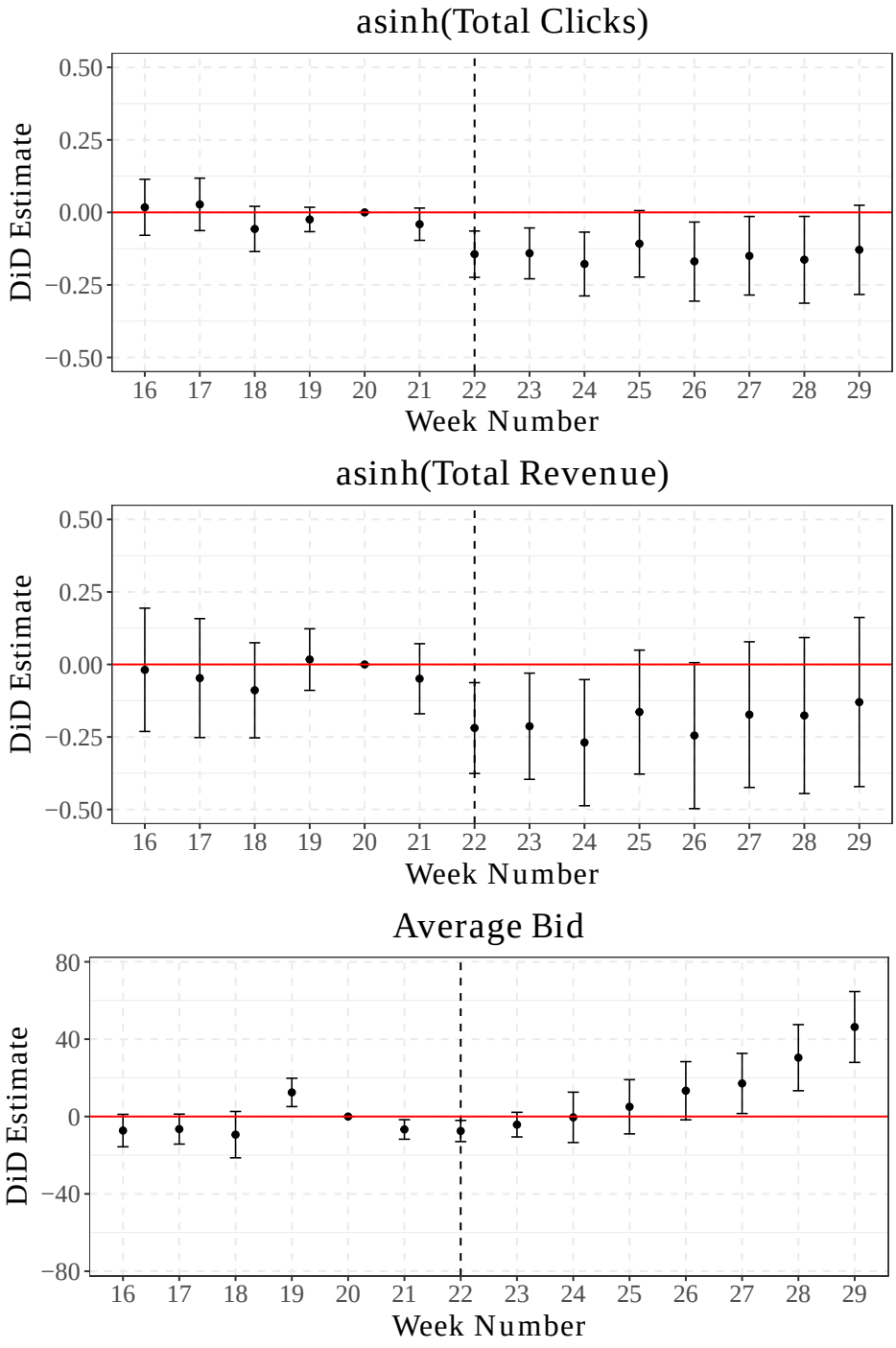


Table 3: Difference-in-Differences Estimates for Advertising Outcome Variables

	(1) asinh(Total Clicks)	(2) Total Clicks	(3) asinh(Distinct Clicks)	(4) Distinct Clicks	(5) asinh(Revenue)	(6) Revenue	(7) Average 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.

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

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

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

Notes: t-statistics are reported in parentheses. The standard errors for every regression are clustered at the website-country level. We aggregate every dependent variable to the website-country-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.

References

Abadie, A., A. Diamond, and J. 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.

Acemoglu, D., A. Makhdoumi, A. Malekian, and A. Ozdaglar (2019). Too much data: Prices and inefficiencies in data markets. Technical report, National Bureau of Economic Research.

Acquisti, A., L. K. John, and G. Loewenstein (2013). What is privacy worth? *The Journal of Legal Studies 42*(2), 249–274.

Acquisti, A., C. Taylor, and L. Wagman (2016). The economics of privacy. *Journal of Economic Literature 54*(2), 442–92.

Aridor, G., Y. Mansour, A. Slivkins, and Z. S. Wu (2020). Competing bandits: The perils of exploration under competition. *arXiv preprint arXiv:2007.10144*.

Athey, S., C. Catalini, and C. Tucker (2017). The digital privacy paradox: Small money, small costs, small talk. Technical report, National Bureau of Economic Research.

Bajari, P., V. Chernozhukov, A. Hortaçsu, and J. Suzuki (2019). The impact of big data on firm performance: An empirical investigation. In *AEA Papers and Proceedings*, Volume 109, pp. 33–37.

Bellemare, M. F. and C. J. Wichman (2019). Elasticities and the inverse hyperbolic sine transformation. *Oxford Bulletin of Economics and Statistics*.

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

Bergemann, D., A. Bonatti, and T. Gan (2022). The economics of social data. *The RAND Journal of Economics*.

Boerman, S. C., S. Kruikemeier, and F. J. Zuiderveen Borgesius (2018). Exploring motivations for online privacy protection behavior: Insights from panel data. *Communication Research*, 0093650218800915.

Braghieri, L. (2019). Targeted advertising and price discrimination in intermediated online mar-

kets. *Available at SSRN 3072692*.

Buchholz, N., L. Doval, J. Kastl, F. Matějka, and T. Salz (2022). The value of time: Evidence from auctioned cab rides. Technical report, National Bureau of Economic Research.

Cameron, A. C. and P. K. Trivedi (1990). Regression-based tests for overdispersion in the poisson model. *Journal of Econometrics 46*(3), 347–364.

Cameron, A. C. and P. K. Trivedi (2005). *Microeconometrics: methods and applications*. Cambridge university press.

Campbell, J., A. Goldfarb, and C. Tucker (2015). Privacy regulation and market structure. *Journal of Economics & Management Strategy 24*(1), 47–73.

Chiou, L. and C. Tucker (2017). Search engines and data retention: Implications for privacy and antitrust. Technical report, National Bureau of Economic Research.

Choi, J. P., D.-S. Jeon, and B.-C. Kim (2019). Privacy and personal data collection with information externalities. *Journal of Public Economics 173*, 113–124.

Coey, D. and M. Bailey (2016). People and cookies: Imperfect treatment assignment in online experiments. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 1103–1111.

Degeling, M., C. Utz, C. Lentzsch, H. Hosseini, F. Schaub, and T. Holz (2018). We value your privacy... now take some cookies: Measuring the gdpr's impact on web privacy. *arXiv preprint arXiv:1808.05096*.

DeGroot, M. H. and S. 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, B. A. and J. J. Harden (2013). Testing for zero inflation in count models: Bias correction for the vuong test. *The Stata Journal 13*(4), 810–835.

DPC (2020a). Guidance note: Cookies and other tracking technologies.

DPC (2020b). Report by the data protection commission on the use of cookies and other tracking technologies.

Dubé, J.-P. and S. Misra (2019). Personalized pricing and customer welfare. *Available at SSRN*

2992257.

Einav, L. and J. Levin (2014). Economics in the age of big data. *Science 346*(6210), 1243089.

Fairfield, J. A. and C. Engel (2015). Privacy as a public good. *Duke LJ 65*, 385.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters 27*(8), 861–874.

Godinho de Matos, M. and I. Adjerid (2021). Consumer consent and firm targeting after gdpr: The case of a large telecom provider. *Management Science*.

Goldberg, S., G. Johnson, and S. Shriver (2021). Regulating privacy online: An economic evaluation of the gdpr. *Available at SSRN 3421731*.

Goldfarb, A. and C. Tucker (2012a). Privacy and innovation. *Innovation policy and the economy 12*(1), 65–90.

Goldfarb, A. and C. Tucker (2012b). Shifts in privacy concerns. *American Economic Review 102*(3), 349–53.

Goldfarb, A. and C. E. Tucker (2011). Privacy regulation and online advertising. *Management Science 57*(1), 57–71.

Jia, J., G. Z. Jin, and L. Wagman (2018). The short-run effects of gdpr on technology venture investment. Technical report, National Bureau of Economic Research.

Jia, J., G. Z. Jin, and L. Wagman (2020). Gdpr and the localness of venture investment. *Available at SSRN 3436535*.

Johnson, G. (2013). The impact of privacy policy on the auction market for online display advertising.

Johnson, G., S. Shriver, and S. Goldberg (2020). Privacy & market concentration: Intended & unintended consequences of the gdpr. *Available at SSRN 3477686*.

Johnson, G. A., S. K. Shriver, and S. Du (2020). Consumer privacy choice in online advertising: Who opts out and at what cost to industry? *Marketing Science*.

Kehoe, P. J., B. J. Larsen, and E. Pastorino (2018). Dynamic competition in the era of big data. Technical report, Working paper, Stanford University and Federal Reserve Bank of Minneapolis.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics 34*(1), 1–14.

Liang, A. and E. Madsen (2019). Data sharing and incentives. *Available at SSRN 3485776*.

Lin, T. (2022). Valuing intrinsic and instrumental preferences for privacy. *Marketing Science*.

Lin, T. and S. Misra (2022). Frontiers: the identity fragmentation bias. *Marketing Science*.

Norberg, P. A., D. R. Horne, and D. A. Horne (2007). The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs 41*(1), 100–126.

Peukert, C., S. Bechtold, M. Batikas, and T. Kretschmer (2022). Regulatory spillovers and data governance: Evidence from the gdpr. *Marketing Science*.

Prince, J. T. and S. Wallsten (2020). How much is privacy worth around the world and across platforms? *Journal of Economics & Management Strategy*.

Utz, C., M. Degeling, S. Fahl, F. Schaub, and T. Holz (2019). (un) informed consent: Studying gdpr consent notices in the field. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 973–990. ACM.

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

Zhuo, R., B. Huffaker, S. Greenstein, et al. (2021). The impact of the general data protection regulation on internet interconnection. *Telecommunications Policy 45*(2), 102083.

Zou, Y., K. Roundy, A. Tamersoy, S. Shintre, J. Roturier, and F. Schaub (2020). Examining the adoption and abandonment of security, privacy, and identity theft protection practices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–15.

Appendix

A Omitted Figures and Tables

Omitted Consumer Response Figures

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

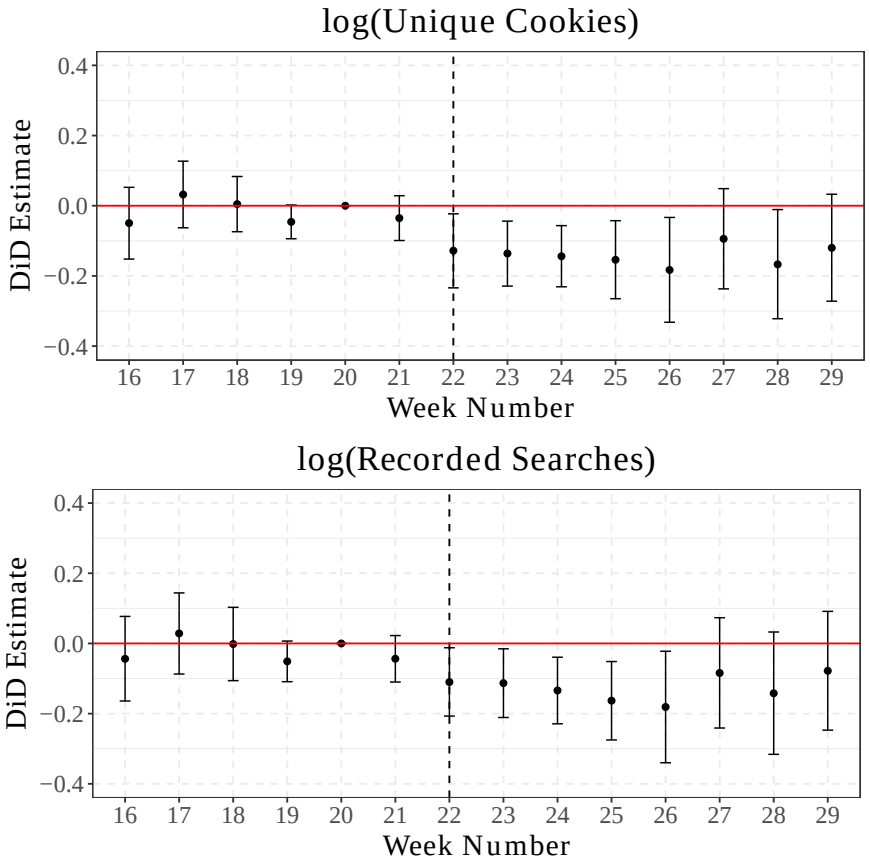


Table 5: Summary Statistics of Consumer Persistence

Region	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)

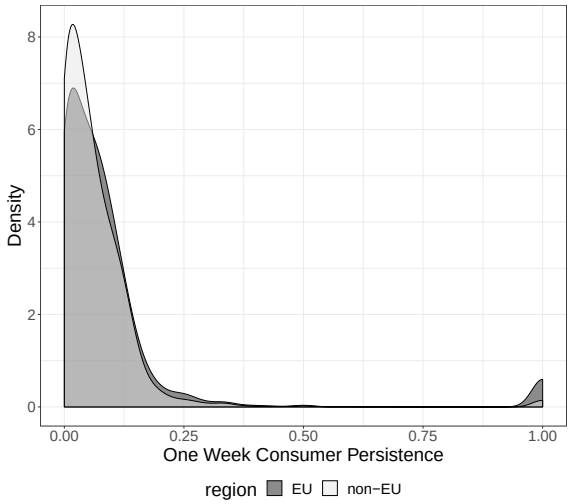


Table 6: Difference-in-Differences Estimates for Sales Activity

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

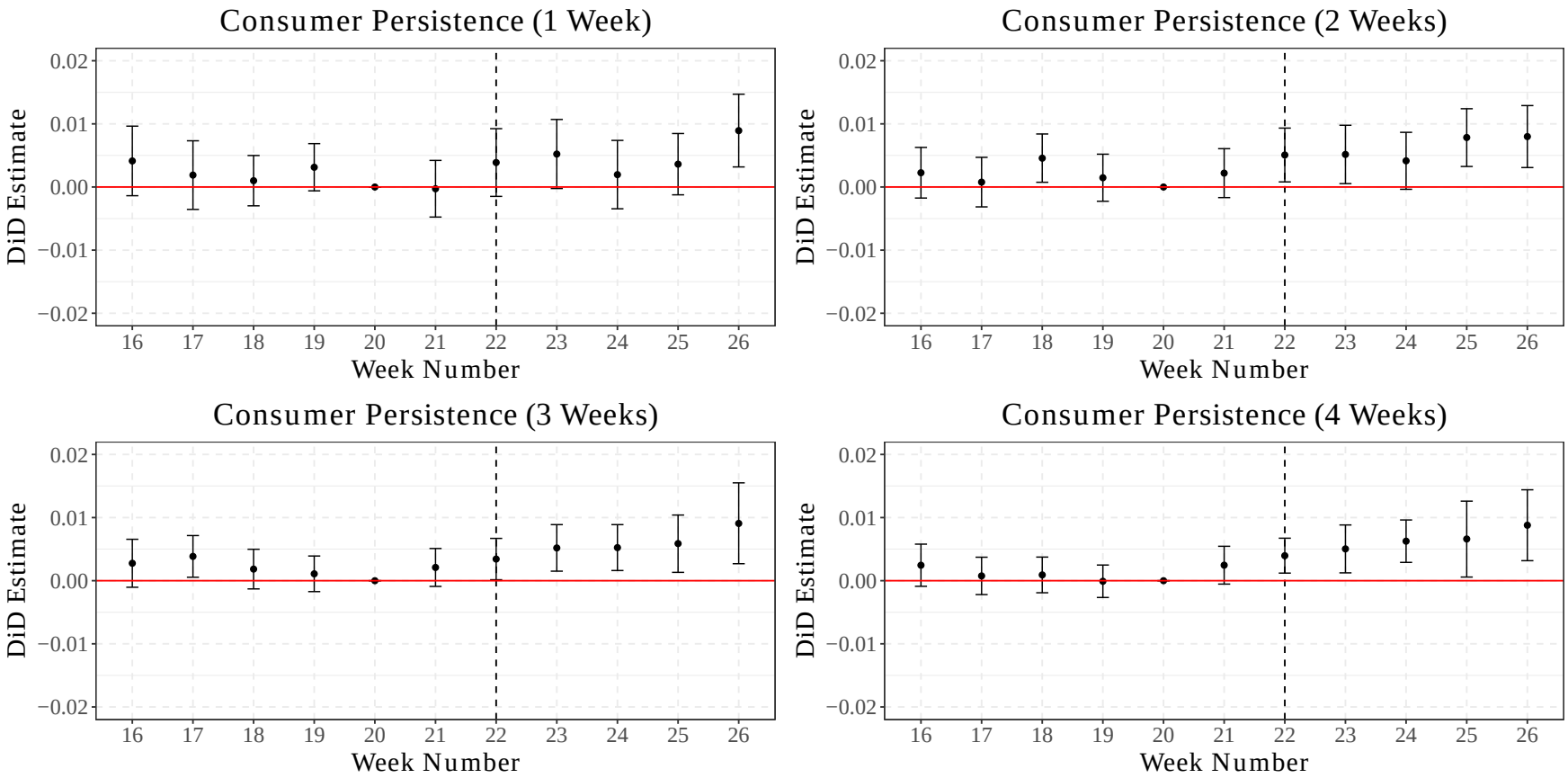


Table 7: Persistence - Large vs. Small Website Heterogeneous Treatment Effects

	(1) 1 Week Persistence	(2) 2 Weeks Persistence	(3) 3 Weeks Persistence	(4) 4 Weeks Persistence
Treated	0.000204	0.00230	0.00183	0.00347*
	(0.10)	(1.29)	(1.21)	(1.85)
Large Website × Treated	0.00367	0.00239	0.00255*	0.00202
	(1.49)	(1.22)	(1.79)	(1.22)
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. We report heterogeneous treatment effects across large vs. small websites. A website is denoted as large (resp. small) if they are above (resp. below) the median website size where size is defined as average number of weekly searches on the website pre-GDPR. The omitted category is small websites.

Omitted Advertising Figures

Table 8: Difference-in-Differences Estimates for Advertisements Delivered

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

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

Notes: t-statistics are reported in parentheses. The standard errors for every regression are clustered at the website-country level. We aggregate every dependent variable to the website-country-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 overall level and inverse hyperbolic sine transform of total advertisements delivered to consumers.

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

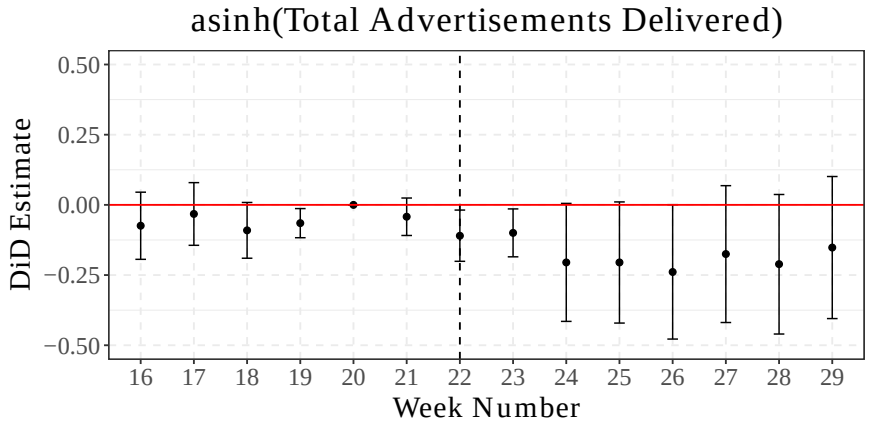


Table 9: Difference-in-Differences Estimates for Fraction of Targeted Advertisements

	(1) Fraction Exposed to Ads Overall	(2) Fraction Exposed to Ads Multi-searchers	(3) Fraction Exposed to Ads Single-searchers
DiD Coefficient	0.0176**	0.0242***	0.0165*
	(2.06)	(3.09)	(1.73)
OS + Browser Controls	✓	✓	✓
Week FE			
Website × Country FE	✓	✓	✓
Observations	3602	3602	3602

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

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-period level between April 13th and July 20th. The period in the aggregation denotes pre vs. post GDPR. We aggregate to the period level in order to have a sufficiently long period to track if someone was ever exposed to an advertisement and is a single-searcher. The dependent variable in column (1) is the fraction of observed cookies that were ever targeted for advertisements. The dependent variable in column (2) is the fraction of observed multi-searcher cookies that were ever targeted for advertisements. The dependent variable in column (3) is the fraction of observed single-searcher cookies that were ever targeted for advertisements.

Table 10: Overall Single vs. Multi Searcher Flight Search Patterns

Days Before Departure		Fraction of Domestic Flights	
Single-Searcher	Multi-Searcher	Single-Searcher	Multi-Searcher
57.5	58.0	0.305	0.297

Notes: The first two columns show the mean number of days before departure for single-searchers and multi-searchers. The final two columns show the mean fraction of domestic flights for single-searchers and multi-searchers. The statistics are computed across the dataset of the website-country-browser-OS-period level aggregation between April 13th - July 20th. The period in the aggregation denotes pre vs. post GDPR.

Table 11: Single vs. Multi Searcher Flight Search Patterns by Region

Region	**Days Before Departure**	Fraction of Domestic Searches
non-EU	**Multi-searchers**: 58.4 (0.57)	**Multi-searchers**: 0.477 (.005)
	Single-searchers: 58.1 (0.39)	**Single-searchers**: 0.474 (.005)
EU	**Multi-searchers**: 57.7 (0.58)	**Multi-searchers**: 0.145 (0.002)
	Single-searchers: 56.9 (0.49)	**Single-searchers**: 0.165 (0.003)

Notes: The table displays the fraction of searches that are for domestic flights before and after the GDPR across EU vs. non-EU countries. The table separately displays the fraction for multi-searchers – cookies associated with more than one search –and single-searchers – cookies associated with only one search – across EU and non-EU countries. Standard errors are in parentheses.

Table 12: Mean Market Concentration

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

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

Table 13: Difference-in-Differences Estimates for Composition of Advertisers

	(1) CR-1	(2) CR-3	(3) CR-5	(4) HHI	(5) Mean Number of Bidders
DiD Coefficient	0.00233*	0.00826**	0.0138**	0.00227*	-.363***
	(1.78)	(2.17)	(2.24)	(1.95)	(-4.38)
OS + Browser Controls	✓	✓	✓	✓	✓
Product Category Controls	✓	✓	✓	✓	✓
Week FE	✓	✓	✓	✓	✓
Website × Country FE	✓	✓	✓	✓	✓
Observations	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 variables in the regressions reported in the first three columns are the market share concentrations of the top 1, top 3, and top 5 advertisers according to share of advertisements delivered to consumers (concentration ratio 1, 3, and 5 respectively). The dependent variable in the fourth column is the Herfindahl-Hirschman Index (HHI) using the same market share definition. The dependent variable in the fifth column is the mean number of bidders in the auction. We define $CR\text{-}I_{jt}$ as the concentration of impressions for the top I out of K advertisers on website j at time t. Let imp_{kjt} be the impressions of the k-th largest advertiser (according to impression share) on website j at time t. Then, $CR\text{-}I_{jt} = \frac{\sum_{k=1}^{I} imp_{kjt}}{\sum_{k=1}^{K} imp_{kjt}}$. For a website j and time t, the share of advertiser $k \in \{1, 2, ..., K\}$ is denoted by s_k. HHI is therefore defined as follows: $HHI_{jt} = \sum_{k=1}^{K} s_k^2$.

Omitted Prediction Figures

Figure 12: Week by Week Treatment Effect (Average Predicted Probability and Class Proportion)

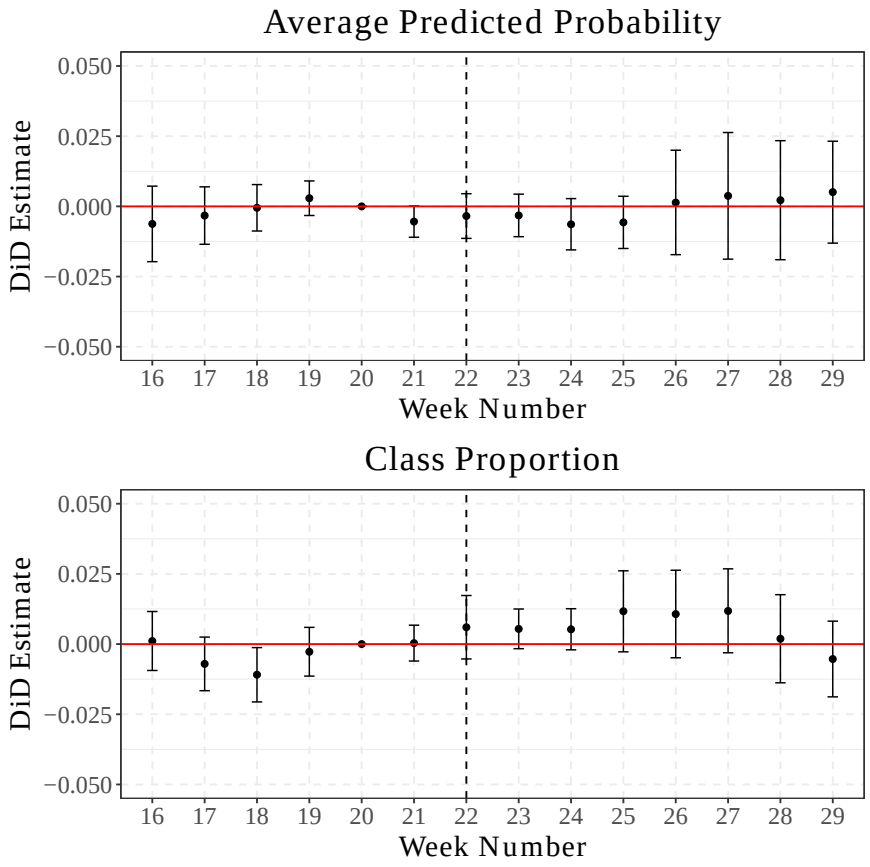
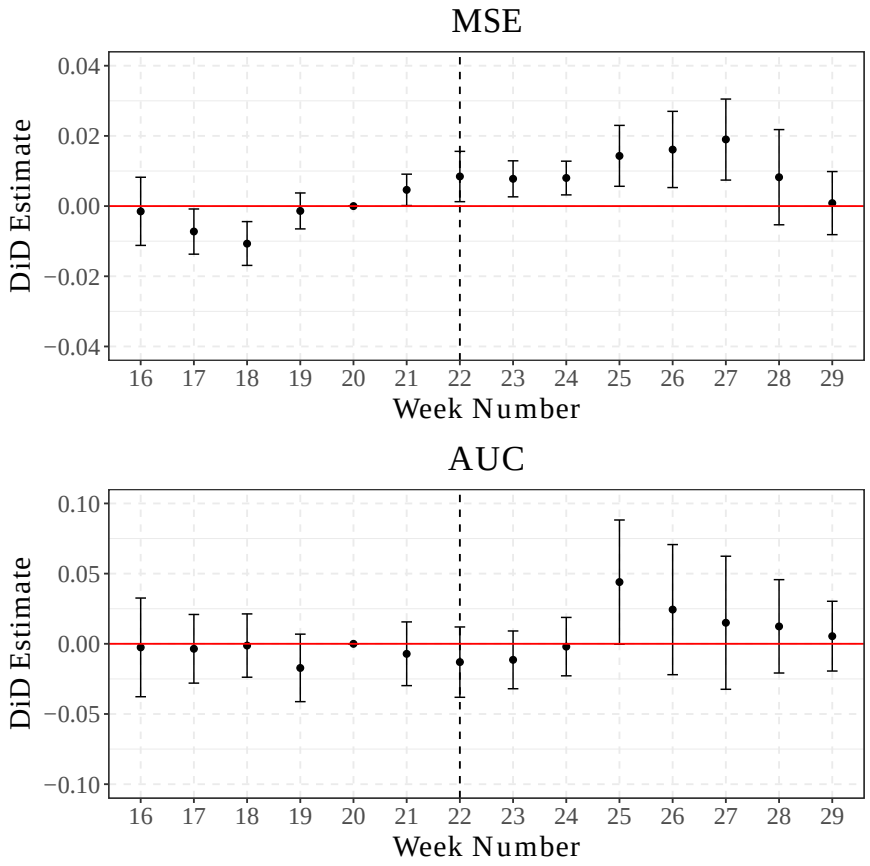


Figure 13: Week by Week Treatment Effect (MSE and AUC)



B Model of Single Searcher Inflation

In this section we more formally investigate how the distribution of searches changes after GDPR in order to better understand the mechanisms driving the increase in persistence observed in section 4. We model the "single searcher" observation from section 4 as individuals obfuscating their identities using advertising/cookie blockers that generate distinct "excess" single searchers relative to the natural, non-obfuscated, distribution of searches. Our main objective is to understand the following. First, we investigate whether consumers were making use of such privacy means prior to the introduction of GDPR by estimating the fraction of "excess" single searchers. Second, consistent with *privacy means substitution*, we determine whether the fraction of consumers doing so has decreased after GDPR is introduced. Third, consistent with *selective consent*, we explore whether the search distribution of the post-GDPR searches has changed so that the remaining consumers search more on the given website.

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 $k \geq 1$. The distribution of observed history length for the obfuscation consumer type o is degenerate with probability mass 1 at $k = 1$. 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:

$$S_1 = \pi(x) + (1 - \pi(x)) \cdot Q(1; \theta(x)) \tag{5}$$

$$S_k = (1 - \pi(x)) \cdot Q(k; \theta(x)), \quad \forall k \geq 1 \tag{6}$$

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 these estimates, we can compute the mean search history length before and after GDPR denoted as $\bar{Q}(\hat{\theta}^{PRE})$ and $\bar{Q}(\hat{\theta}^{POST})$ respectively. Given this setup, our informal hypotheses can be stated as the following null hypotheses:

1. $H_0 : \overline{\hat{\pi}^{PRE}} = 0$, $H_a : \overline{\hat{\pi}^{PRE}} \neq 0$
2. $H_0 : \overline{\hat{\pi}^{POST}} = \overline{\hat{\pi}^{PRE}}$, $H_a : \overline{\hat{\pi}^{POST}} < \overline{\hat{\pi}^{PRE}}$
3. $H_0 : \bar{Q}(\hat{\theta}^{POST}) = \bar{Q}(\hat{\theta}^{PRE})$, $H_a : \bar{Q}(\hat{\theta}^{POST}) > \bar{Q}(\hat{\theta}^{PRE})$

Data and Estimation

For this exercise we restrict attention to the same large hotels website shown in Figure 6. 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 $\mu(x)$, $\alpha(x)$.[44]

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 to evaluate whether a model with type o consumers leads to a better fit to the observed data (Vuong, 1989; Desmarais and Harden, 2013). In order to test our second and third hypotheses of interest, we do a t-test comparing the vectors of $\hat{\pi}^{PRE}$ and $\hat{\pi}^{POST}$ as well as $Q(\hat{\theta}^{PRE})$ and $Q(\hat{\theta}^{POST})$ respectively.

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 $\overline{\hat{\pi}^{POST}} \geq \overline{\hat{\pi}^{PRE}}$. We are able to reject the null with $p < 0.001$. 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. Furthermore, we compare the resulting estimates of $Q(\hat{\theta}^{PRE})$ and $Q(\hat{\theta}^{POST})$ and run a t-test with the null hypothesis that $\bar{Q}(\hat{\theta}^{POST}) = \bar{Q}(\hat{\theta}^{PRE})$. We are able to reject the null with $p < 0.001$ and with a mean increase of nearly one search per consumer as $\bar{Q}(\hat{\theta}^{PRE}) = 4.37$ and $\bar{Q}(\hat{\theta}^{PRE}) = 5.30$. Thus, this provides evidence that both the fraction of obfuscators fell, but also

[44] 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.

that the non-obfuscators search more.

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 ($p < 0.001$) 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).

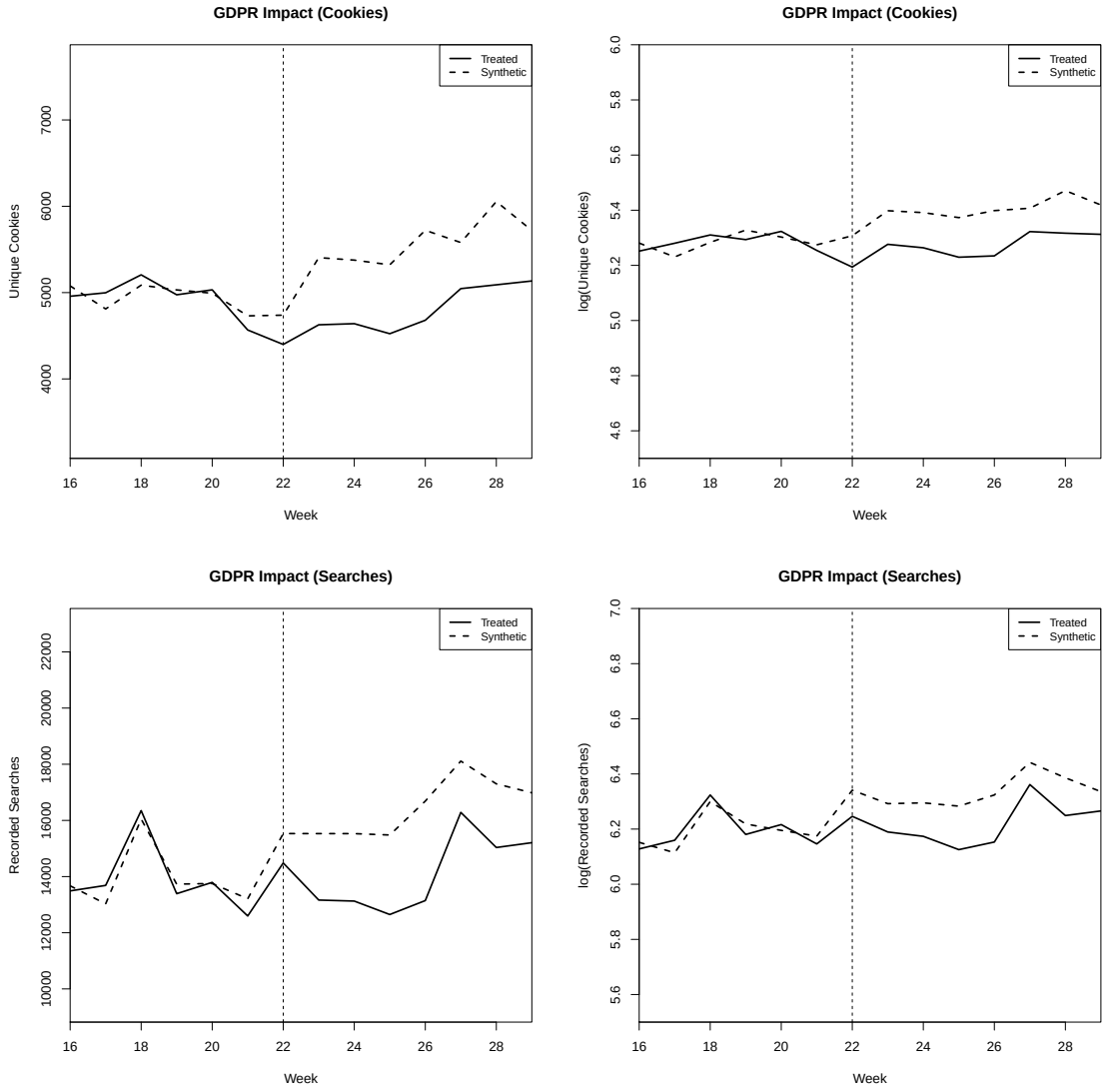
Under the the assumption of a negative binomial, the Vuong test still concludes that there is evidence for the presence of type o consumers ($z = -6.97$) in the pre-GDPR period. However, we no longer reject the model without excess single searchers in the post-GDPR period ($z = -1.81$, AIC-corrected: $z = -0.84$, BIC-corrected: $z = 4.94$). Furthermore, we are able again to reject the null hypothesis that $\hat{\pi}^{POST} = \hat{\pi}^{PRE}$ and $\bar{Q}(\hat{\theta}^{POST}) = \bar{Q}(\hat{\theta}^{PRE})$ with $p < 0.001$.

In sum, we document statistical evidence for obfuscators 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. Finally, under both distributional assumptions, we find that the mean of the non-obfuscator search history length distribution has increased post-GDPR.

Online Appendix

A Robustness for Consumer Response Results

Figure A1: 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 log of the number of unique cookies (top) and the log of the 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 A.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 A.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.

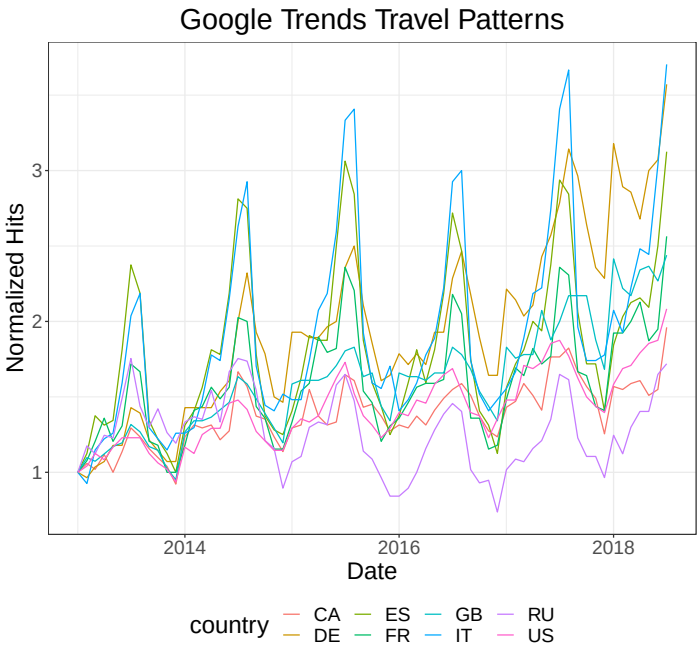
A.1 Synthetic Controls

In order to provide additional validation for the difference-in-differences results in section 4, we supplement our primary analysis with a synthetic controls analysis, following Abadie et al. (2010) and utilizing the corresponding R package, `Synth`. We aggregate the data to the same level as we do in the primary analysis.[1] 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.[2] 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 A1. 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.

[1] 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.

[2] 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 we 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 A2: 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.

A.2 Controlling for Differences in Travel Patterns Across Countries

Since our article 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 result in our estimates understating 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 impact of these controls on our estimates of the change in total recorded searches and unique

cookies.

Table A1: 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

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

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.

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.[3]

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

[3] Google Trends Documentation (Retrieved on May 1st, 2022) provides additional details.

the countries in our analysis: booking. Figure A2 plots the results from Google Trends for the trends for this keyword from January 1st, 2013 until July 31st, 2018. Figure A2 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 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:[4]

$$\text{google}_{c,t} = \chi\Big[week \times country\Big] + \epsilon_{c,t} \tag{7}$$

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

$$y_{t,c,j,o,b,p} = \alpha_t + \hat{\chi}_{c,t} + \delta_{j,c} + \gamma_o + \zeta_b + \omega_p + \beta(EU_j \times after_t) + \epsilon_{t,c,j,o,b,p} \tag{8}$$

where the notation is identical to that utilized in the main text and $\hat{\chi}_{c,t}$ denotes the coefficient on $week \times country$ that comes from from running (7). The regression results are reported in Table A1 and are quantitatively consistent with the results from our main specification.

B Persistence Treatment Effects by Browser/OS

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 cookie management.

[4]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.

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:

$$y_{t,c,j,o,b,p} = \alpha_t + \delta_{j,c} + \gamma_o + \zeta_b + \omega_p + \beta(EU_j \times after_t \times browser) + \epsilon_{t,c,j,o,b,p} \tag{9}$$

We focus primarily on browsers where it is clear that the two mechanisms should lead to different conclusions. For instance, if we observed similar persistence across Internet Explorer and Google Chrome, this would provide strong evidence for selective consent over privacy means substitution. Internet Explorer is a web browser primarily used on older computers and is known to attract older, less technologically sophisticated users since, at the time of our analysis, it was already phased out by Microsoft for Edge on newer Windows computers. Furthermore, the lack of JavaScript extensions on Internet Explorer makes cookie blockers substantially more difficult to implement on the browser 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.[5] We may further expect that different levels of privacy protection among browsers would induce selection in browser choice based on privacy preferences. For instance, Apple Safari and Mozilla Firefox at the time of the GDPR had a broad set of privacy protection means built into it, whereas Google Chrome had laxer privacy controls.[6] Although the selection into web browsers based on privacy preferences, especially during the time period of analysis, is not as clear cut of an indicator, if we observe little differences between these browsers, then it provides some evidence for selective consent.

Table A3 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 privacy means substitution. The estimated treatment effect is lower in Safari relative to Chrome, but the difference is not sta-

[5] See, for instance, Is Ad Block Available on Internet Explorer? Retrieved on January 29th, 2022.

[6] 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.

tistically significant. Both of these observations are consistent with privacy means substitution, though the lack of definitive differences between Chrome and Safari/Firefox does suggest that selective consent may also play a role.

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:

$$y_{t,c,j,o,b,p} = \alpha_t + \delta_{j,c} + \gamma_o + \zeta_b + \omega_p + \beta(EU_j \times after_t \times OS) + \epsilon_{t,c,j,o,b,p} \tag{10}$$

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 privacy means substitution, we would expect a larger difference in persistence on desktop compared to mobile whereas for consistency with selective consent we should expect a smaller difference.

Table A2 displays the regression results with Windows as the omitted operating system. We focus on the point estimates for the heterogeneous effects for Android and iOS separately. The estimates for Android indicate a negative, statistically significant effect relative to Windows and, consequently, no increase in persistence for $k = 1, 2$. However, there is no statistically detectable difference between Android and Windows for $k = 3, 4$. For iOS, there is a negative point estimate for $k = 1, 2$ relative to Windows, but this is statistically insignificant. Otherwise, the treatment effect is approximately the same across the different operating systems. We interpret this as a weak difference between persistence on mobile and desktop, which is suggestive of privacy mean substitution, but does not provide conclusive evidence. Overall, the results for both the specifications points to evidence for privacy means substitution, although leaving room for selective consent to play a role in explaining the increase in persistence.

Table A2: Consumer Persistence - 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 A3: 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.

C Prediction Evaluation Measures

C.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 $\hat{p}_{i,j,k}$ is classified as a purchaser if $\hat{p}_{i,j,k} > \hat{P}$ and a non-purchaser if $\hat{p}_{i,j,k} \leq \hat{P}$. This would result in a false positive rate—a rate at which a non-purchaser is misclassified into a purchaser:

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

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

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

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

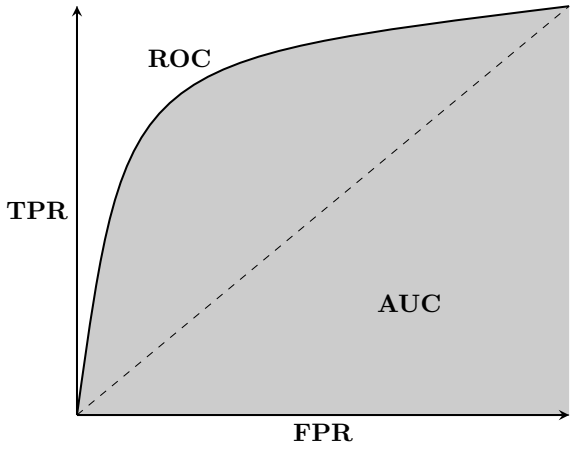
The ROC is obtained by tracing the locus of (FPR, TPR) by varying the classification threshold $\hat{P}$.[7] The slope of the ROC corresponds to the additional *power* (in rate) the prediction gains for an additional unit of type I error (in rate) it tolerates. For a random predictor, this slope would 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 A3 depicts a typical ROC curve.

C.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

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

Figure A3: 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.

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.[8] For notation, p_k denotes the kth estimated probability bin, n_k denotes the number of probability estimates falling into the kth bin and $\bar{o}_k$

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

denotes the true class proportion in the kth bin in the data. This allows us to rewrite (4) as:

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

Table A4: Difference-in-Differences Estimates for Relevance and Calibration

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

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

Notes: t-statistics are reported in parentheses. The standard errors for every regression are clustered at the website-country level. We aggregate every dependent variable to the website-country-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.

We run the same specification utilizing each component of the decomposition of the MSE as the outcome variable. These results are reported in Table A4. 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 because this increase is a result of the increased uncertainty

in the predicted class (i.e. the class proportion moving closer to 0.5).

D Impact of Persistence and Data Size 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, $pred_{t,c,j,p}$ 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:

$$pred_{t,c,j,p} = \beta \cdot \log(\text{Recorded Searches}) + \alpha_t + \delta_{j,c} + \omega_p + \epsilon_{t,c,j,p} \tag{12}$$

$$pred_{t,c,j,p} = \beta \cdot \text{Persistence} + \alpha_t + \delta_{j,c} + \omega_p + \epsilon_{t,c,j,p} \tag{13}$$

Table A6 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.[9] This corresponds to the regression results in Columns (1) and (3) of Table A6. 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.[10]

Table A5 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 A5, 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 increase in consumer persistence should marginally improve prediction according to both AUC and MSE. However, this does not imply that the scale of data is unimportant which would run counter to standard statistical intuition; on the contrary, prediction ability improves substantially as the scale of data increases. Rather, the change in the scale of the data as a result of GDPR is not large enough to cause meaningful changes in prediction error in the long run. However, the increase in persistence as a result of GDPR should lead to an improvement in prediction capabilities in the long run.

[9] 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.

[10] 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 A7 shows the results for the same specification, but uses a sliding window total of recorded searches instead of the weekly total number of recorded searches, and shows that the point estimates do not change much when taking this into account.

Table A5: 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.

Table A6: 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 A7: 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.