Assessing the Value of Internet Data for Medical Applications

Gilie Gefen
Technion
gilie@campus.technion.ac.il

Omer Ben-Porat
Technion
omerbp@campus.technion.ac.il

Moshe Tennenholtz
Technion
moshet@ie.technion.ac.il

Elad Yom-Tov
Microsoft Research Israel and Technion
eladyt@microsoft.com

Abstract

People increasingly turn to the Internet when they have a medical condition. The data they create during this process is a valuable source for medical research and future health services. However, utilizing these data could come at a cost to user privacy. Thus, it is essential to balance the perceived value that users assign to these data with the value of the services derived from them.

Here we describe experiments where methods from Mechanism Design were used to elicit truthful valuations from users for their Internet data and for services to screen people for medical conditions. In these experiments, 880 people from around the world were asked to participate in an auction to provide their data for uses differing in their contribution to the participant, to society, and in the disease they addressed. Some users were offered monetary compensation for their participation, while others were asked to pay to participate.

Our findings show that 99% of people were willing to contribute their data in exchange for monetary compensation and an analysis of their data, while 53% were willing to pay to have their data analyzed. The average perceived value users assigned to their data was estimated at US$49. Their value to screen them for a specific cancer was US$22, while the value of this service offered to the general public was US$20. Participants requested higher compensation when notified that their data would be used to analyze a more severe condition. They were willing to pay more to have their data analyzed when the condition was more severe when they had higher education or if they had recently experienced a serious medical condition.

Our findings show that it is possible to place a monetary value on health-related uses of highly personal data. Such uses are valued by users, and their value is approximately half that of their data. Our methodology can be extended to other areas where sensitive data may be exchanged for services to individuals and to society, while our results suggest that future services utilizing individual's Internet data could be viable.

1 Introduction

Data is a valuable asset for organizations in a data-driven economy (Dewar 2017), but data holders can abuse this asset, one result of which is the possible breach of an individual's privacy. In response to the latter and to other issues stemming from amassing data by companies, the European Union's General Data Protection Regulation (GDPR) which came into effect on May 2018 gives control to individuals over their personal data (eugdpr.org), thus attempting to shift the balance between the value of data for individuals and for enterprises.

Though it is evident that data has value, quantifying it is difficult. In the past, researchers have attempted to put a monetary value that individuals assign to their photos (Carrascal and de Oliveira 2013), browsing (Carrascal and de Oliveira 2013), data from home appliances (Kugler 2018) and location (Danezis, Lewis, and Anderson 2005). Additionally, researchers have proposed "active choice" models which offer users the option of payment through monetary transactions or disclosure of personal information (Malgieri and Custers 2018).

One area where the value of the data has not been measured is Healthcare. It is a significant lacuna because Healthcare is, arguably, the area in which data has the highest value to individuals. On the one hand, the high value is due to the potential for damage if privacy is breached, and on the other hand, for companies and individuals, because of the potential uses for creating new screening, treatment, and insights from data.

In recent years data from online services has proven a boon for medical research. For example, the website PatientsLikeMe (www.patientslikeme.com) lets people who are suffering from one of several diseases connect with others that suffer from similar problems for social support and advice. The data posted by users of this site has enabled researchers to test the efficacy of treatments, which would have been challenging to test through other methods, for example, in rare diseases such as ALS (Wicks et al. 2011).

Another source of data from online services is people's queries on Internet search engines. These queries and the interactions of users with these search engines have been used to screen for severe medical conditions such as Parkinson's disease (Allerhand et al. 2018; White, Doraiswamy, and Horvitz 2018), ovarian and cervical cancer (Soldaini and Yom-Tov 2017), lung cancer (White and Horvitz 2017), pancreatic cancer (Paparrizos, White, and Horvitz 2016) and diabetes (Hochberg et al. 2019). Similarly, social media postings have been used to predict depression (De Choudhury

et al. 2013) and diagnose autism (Ben-Sasson and Yom-Tov 2016).

Thus, people's data in online services can be used to create new medical services, e.g., screening tests for disease, but these data could potentially compromise individual privacy. Therefore, it is essential to assess the value people assign to these data, vis-a-vis the value they perceive to be gained from such novel services.

We note that the value of these services need not be limited to direct monetary value to the individual. People are often willing to donate money, time, effort, data (Skatova, Ng, and Goulding 2014), and even organs (e.g., blood and kidneys) to help others in need, presumably in exchange for societal benefit, personal satisfaction, or to improve their social standing (Misje et al. 2005). Thus, the value of data needs to account for many facets of its perceived value to people and to incorporate the potential harm which might be caused by mishandling thereof.

In this paper, we attempt to measure the value people assign to their search queries on Google or Bing (henceforth search logs) for medical uses.

Measuring the value of data is a difficult undertaking because it requires eliciting a truthful value from people. People who are asked to provide unverifiable information may misreport their valuations if they are skewed towards privacy, without incurring any cost. Thus, here we applied Mechanism Design (Nisan and Ronen 2001) approaches, utilizing two forms of auction, a reverse Vickrey auction in one population and a Vickrey auction in another. In the reverse Vickrey auction, we treat search logs as goods for sale, and design an auction with multiple sellers (agents) and one buyer (the authors of the paper). The buyer informs sellers that he is interested in at most X units of the goods, and is willing to pay at most r dollars for each unit; r is known as the reserve price, and is possibly hidden. Each seller declares her bid (the required price for selling the goods), and the buyer buys X units with the lowest bids, assuming the bids are below the reserve price and pays the minimum between r and the $X + 1$ lowest bid.

Similarly, in the Vickrey auction, we ask participants to provide their search logs and demand that they pay the experimenters to analyze them. In this case, the authors are sellers of an analysis service willing to process at most X units of the goods (search logs), and are willing to accept the X units offered at the highest price, if it is over r, a minimal (possibly hidden) reserve price. The details of the experiment and its modeling are described in Section 2.

The class of these mechanisms is *dominant strategy incentive compatible* (Vickrey 1961), meaning that it is in a participant's best interest to bid her real value for the goods, regardless of X, r or the bids declared by the other sellers. In our study, by letting participants believe that we are willing to pay for their information, we are guaranteed, theoretically, that the elicited values are truthful. Importantly, we debriefed the participants after the experiment, and that too led to interesting findings (see Section 4).

Our paper provides several contributions. First, we develop a novel method for measuring and modeling the monetary value of data and data services. Second, we show that people assign a high value to health data and health-related services. The value of the latter is such that almost half the people are willing to pay for these services, even when these services are provided to the public, without direct benefit to themselves. Our models suggest that the perceived value of data is approximately twice as that of the proposed health service using these data.

2 Methods

2.1 Data collection

We conducted an online assessment of the value people assign to their internet search history data for its use in medical purposes by running an auction/reverse auction, extracting people's valuations for health services and data in 8 different conditions. Participants were recruited to participate in an online questionnaire through two crowdsourcing platforms, Mechanical Turk and Prolific Academic. The two platforms differ in workers' demographics employed in each, their geographic reach, and the attention of workers to the task (Peer et al. 2017). For participating in the study, the participants who completed the questionnaire received US$1.50.

At the beginning of the questionnaire, we provided participants with background information, stating that search logs have been used for medical purposes.

We then requested participants to consent to participate in the study and to provide their birth year, gender, country of residence, level of education, and yearly income. They were also required to indicate if they have recently suffered from a severe medical condition and/or are currently suffering from it.

We informed participants of the **stated goal** of the questionnaire, which was to screen participants for further study in which those chosen would provide their search history on Bing and Google so that experimenters could use it towards one of 4 goals:

1. **Benign medical condition, public good**: Measure and report the rate of flu virus in the participant's country.
2. **Severe medical condition, public good**: Develop a model to detect thyroid cancer using their data and apply it to people in the participant's country.
3. **Severe medical condition, personal good**: Apply a model for the detection of thyroid cancer to the participant's data and report the result to them.
4. **Severe medical condition, public and personal good**: Develop a model to detect thyroid cancer, apply it to the participant's data, and report the result to them.

Participants were randomized into **one** of the 4 goals, as well as to one of two additional conditions: Either the participant was asked how much money (in US dollars) they would require the experimenters to pay for their search history to be used for the stated goal or how much money participants would pay the experimenters to use their search history for the stated goal.

Thus, each participant was randomized into one of 8 experimental conditions (4 goals, pay, or be paid).

As noted above, our goal was to elicit truthful evaluations from participants. Therefore, to check the willingness to pay,

in the experiment, we used a Vickrey auction (Vickrey 1961) with a hidden reserve price. In this auction, the k highest bidders who submitted a bid higher than a reserve price r, win, and pay the maximum between the $k+1$ highest bid and r. In our case, r is hidden.

For checking the desire to be paid, in the experiment, we used a reverse Vickrey auction with a hidden reserve price. In this auction, the k lowest bidders who submitted a bid lower than a reserve price r, win, and are paid the minimum between the $k+1$ lowest bid and r. Both auctions are truthful, regardless of the reserve price. In both auctions, we used reserve prices so that no winners will be selected; this is obtained by having an exorbitantly high reserve price (say \$10K) in the first condition and a very low reserve price (e.g., some negative number) in the second one. While hidden reserve prices are common practice, we also included a description of the way they have been chosen in a debrief to the participants, which was well accepted, as described below.

Thus, participants were informed that the study, conducted by Microsoft, was a first part of a two-staged study. In the first part, the questionnaire would be shown to 1000 people. Among those, the 100 people who requested the smallest amount of money for their data, if it were below a maximal threshold (or offered the largest amount if it were above a minimum threshold, in the second condition) would be contacted for the second stage of the study.

Finally, because the study might be perceived as involving deception, we indicated to participants that we would be providing a debriefing on the study after they completed the questionnaire. Participants who indicated their interest in receiving the debriefing and got paid US\$0.10 for it were provided with the debriefing several weeks after the study was completed.

The authors' Institutional Review Board approved this study. See Section 5 for the full questionnaire.

Participants who read the consent form but did not want to take part in the study, and participants who did not complete the questionnaire were removed from the data.

2.2 Modeling

Let V^i_{pu} and V^i_{pr} be the valuation perceived by participant i to the goods (public or personal, respectively), and let $V^i_S = V^i_{pu} + V^i_{pr}$ be the total value perceived by participant i to the service. Let V^i_D denote the cost for an agent in revealing his valuation (which may have privacy implications but also other ramifications). Let $P^i = V^i_S - V^i_D$ be the total valuation of participant i. Notice that P^i is the value to be reported if/when he/she participates in a truth-revealing auction. Notice that under the above terminology an equation of the form $P_i = V^i_S - V^i_D$ holds for each participant i; in the experiment in which participants are offered payments the reported bid $-P^i$ would be non-negative, and in the experiment in which participants are asked to pay the reported bid P^i would be non-negative. The additive multi-attribute structure of utility we use is classical in economics, and the characterization of conditions justifying it appears already in classical studies (Debreu 1959).

Under the assumption that V_D, V_{pr} and V_{pu} are independent and that P is a linear combination thereof, all proposed transactions (questionnaire responses) can be jointly represented using a linear model where the dependent attribute is P and the independent values are indicators of whether the service was offered in the transaction. Since all transactions involved data, V_D is present in all transactions. Specifically, each transaction is in the form of $P^i = V^i_{pu}X^i_{pu} + V^i_{pr}X^i_{pr} - V^i_D$, where X^i_{pu} and X^i_{pr} are indicators determined by the condition that user was shown.

The linear coefficients (V^i_{pu} and V^i_{pr}) obtained as a solution to the equations above represent the average population valuation for the services and the search log. More specifically, the obtained coefficients can be estimated through linear regression, and the bias term will thus correspond to the average valuation for the revelation of the search logs.

When modeling our data to understand demographic correlates therein and to account for the skewed distribution of the requested and offered amounts, we transformed the non-zero amounts using a log transform before modeling them. Participant's level of education was transformed into a continuous number based on the number of years needed to attain each level of education. For example, participants that indicated high-school as their higher education level received the number "12". Income level was transformed into the average amount in the range, e.g., participants that indicated US\$15,000–US\$30,000 as their yearly income level received the number US\$22,500.

3 Results

We recruited 482 participants through Mechanical Turk and 398 through Prolific Academic. Participants were successfully randomized into one of eight experimental conditions (chi^2 test, $P = 0.47$).

The average reported age of participants was 35 (s.d.: 11) years. The reported gender of 46% was female, 54% male (Less than 1% did not provide it). Participants reported an average of 15 years of education (s.d.: 2). Education was correlated with age (Spearman $\rho = 0.14$, $P = 5 \cdot 10^{-5}$) and with income (Spearman $\rho = 0.27$, $P < 10^{-10}$).

Participants from Mechanical Turk were predominantly from the US (85%) and India (12%), whereas those from Prolific Academic were recruited from 23 countries, the most common being US (28%), UK (23%), and Poland (9%).

Figures 1, 2 and 3 compare the distribution of age, education, and income of participants, respectively, stratified by whether the participants were recruited from Mechanical Turk or Prolific Academic. As the figures show, Prolific Academic participants were typically younger, with fewer years of education and more were in the lower-income bracket.

We excluded 20 participants who responded in under 10 seconds to the monetary value question and another 41 participants who offered or requested more than US\$1500 (9 offered a higher amount and 32 requested a higher amount). Thus, 819 responses (93%) were analyzed. Among 419 participants who were asked to pay, 224 (53%) were willing to pay more than US\$0.10. Similarly, among 400 participants

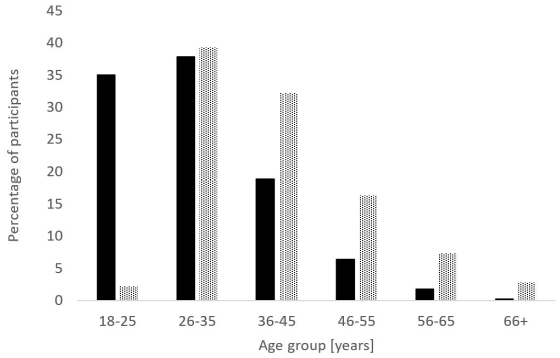

Figure 1: Comparison of the ages of participants, by recruitment platform. Prolific Academic users are shown in black bars, and those from Mechanical Turk in gray.

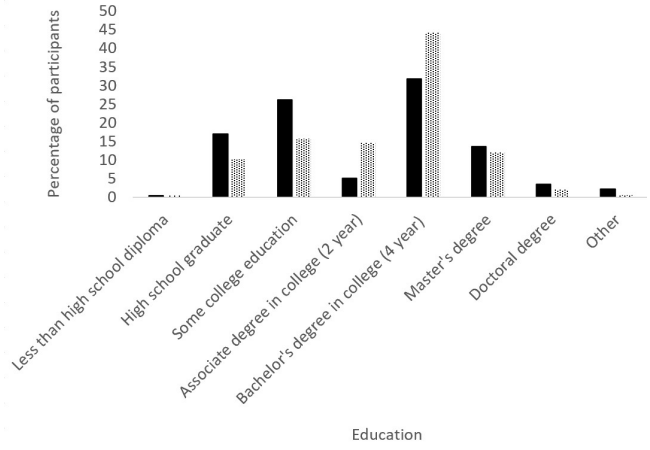

Figure 2: Comparison of the education level of participants, by recruitment platform. Prolific Academic users are shown in black bars, and those from Mechanical Turk in gray.

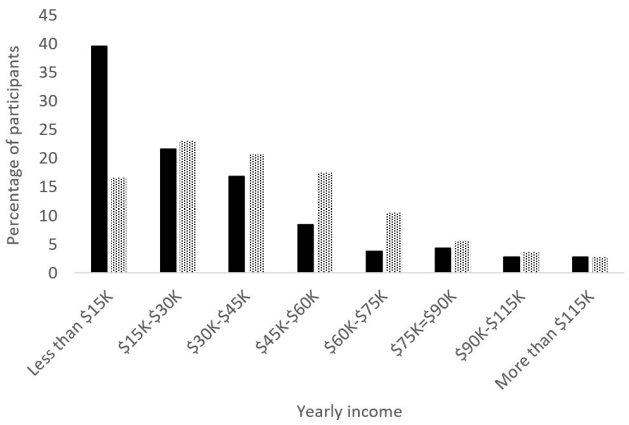

Figure 3: Comparison of reported income, by recruitment platform. Prolific Academic users are shown in black bars, and those from Mechanical Turk in gray.

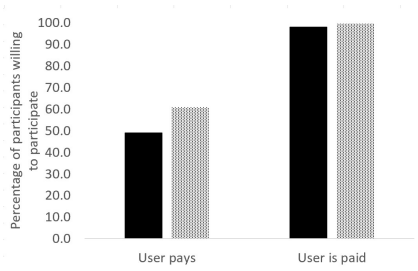

Figure 4: Willingness to provide a monetary value, defined as a willingness to bid a non-zero value. Prolific Academic users are shown in black bars, and those from Mechanical Turk in gray.

who were offered money, 395 (99%) asked for more than US$0.10. There were only small differences in willingness to offer a non-zero value between participants from Mechanical Turk and Prolific Academic, as shown in Figure 4.

The 224 users who offered a payment less than US$0.10 and the five users who requested less than US$0.10 should be considered as **censored users**. Censoring (Dodge 2003) is the failure to observe a variable totally; its value is replaced by a lower limit (right censoring), or an upper limit (left censoring). In our case, a user who offered to pay zero dollars might have been willing to provide their data had she been offered payment, while a user to whom we offered payment and requested a zero amount might have been willing to pay money for their data. However, since we only asked each participant whether they were willing to pay or be paid, our measurements are censored.

A logistic regression model did not find any of the demographic variables statistically significantly associated with being censored. Henceforth we removed the censored users in our analysis. The average value users offered to give for analysis of their data was US$38 (s.d.: 150), and the average value requested was US$148 (s.d.: 269). Figure 5 shows the average bid values per question. As the figure shows, people request a larger amount than they are willing to pay for the service. Additionally, a larger amount is requested for the more severe conditions, but is not offered for such conditions. Strikingly, though the monetary value of public versus personal good is similar when both are proposed, people requested significantly less money and offered slightly less money.

As described in Section 2, it is possible to estimate the average population valuation of the data and the two services by solving a linear regression problem. Thus, we modeled the untransformed value of P using robust linear regression, with 1% of the largest outliers removed. The model is shown in Table 1. As the table shows, the monetary value people attribute to their data is approximately $49. Personal and public goods are valued at approximately $21. The severity of a condition is not statistically significantly correlated with the value offered by participants.

We modeled the monetary value of information, separately for participants who offered payment and those who

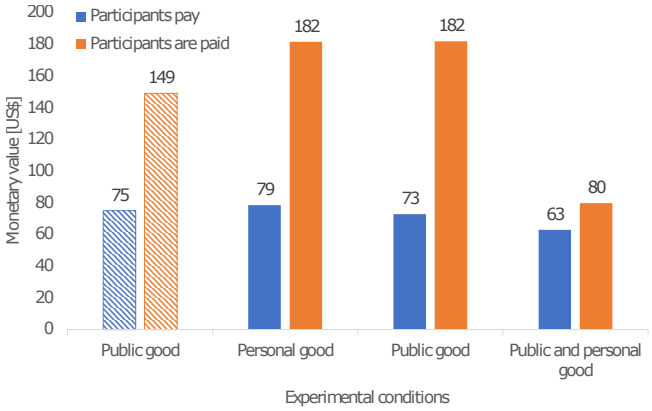

Figure 5: Average bid values after removing censored users. Orange bars represent cases where participants were offered payment and blue bars cases where they were asked to pay. The left two bars are for the benign medical condition, while the other bars refer to the severe medical condition.

Table 1: Model coefficients for predicting monetary value. Only non-censored observations are used (n=619). Model fit is $R^2 = 0.25$.

Variable	Coefficients (SE)	P-value
Value of data	48.8 (12.4)	0.0002
Severity	-13.9 (10.1)	0.174
Personal	22.1 (10.1)	0.028
Public	20.4 (10.1)	0.042

were offered it, excluding censored users. For each group, we constructed one model using only the experimental conditions (severity and type of good) and another which uses both these variables and demographic characteristics reported by users.

Table 2 shows the model coefficients. As the table shows, users were willing to pay more for the use of their data if they had a higher education and if they had experienced a severe medical condition recently. People requested a higher compensation for their data if it were to be used to analyze the more severe condition, and reduced their demand if it contributed to both personal and public good.

Note that in the model, we included the severity as a variable, since it was part of the transaction. However, its $p - value$ was not statistically significant, indicating that it had a negligible effect on the average population valuation.

As noted in Section 2, we offered participants a (paid) debriefing four weeks after they completed the questionnaire. Of 450 people who participated through Mechanical Turk, 363 (81%) asked to receive the debrief, and 353 of 369 (96%) Prolific participants asked for it. Because of the way that the debrief was provided, we could measure how many Prolific participants read the debrief. We found that 326 (92%) read it.

A few participants emailed us after the debrief. Notably, one wrote that "[the study] totally had me fooled!". This would suggest that our methodology of a two-staged study to elicit truthful responses, was effective.

4 Discussion

People increasingly turn to the Internet when they have a medical condition, to diagnose it, learn about their options, and meet other people experiencing similar conditions (Yom-Tov 2016). The data they create during this process is a valuable source for medical research and future health services.

Here we show that when offered compensation, people demand a high price for utilization of their data for health-related services ($149 to $182, depending on condition). The price is higher for use in screening for a severe medical condition and is equal whether the data is used for personal or public good. The latter can be explained by participants perceiving the value of the offered service (apply an algorithm for thyroid cancer detection to their data) as mostly useful for research, not for the individual. Interestingly, when both services are offered, people request less than half the price of each separately.

More than half (53%) of participants were willing to pay the experimenters, in addition to providing their data. Higher value was offered for a service examining the severe medical condition and to personal good (over public good). Strikingly, a lower payment was offered if both goods were satisfied. Higher education and a recent experience of a serious medical condition were associated with higher payment.

When jointly modeled through a linear model described in Section 2, the value of public and personal good (V_{pu} and V_{pr}) were found to be similar, at approximately US$21. In contrast, the perceived value of search logs was estimated at approximately US$49. Thus, the value of search logs (V_D) for health uses (as outlined in the questionnaire) is significantly higher than the value of search logs for other uses, as estimated in a previous study (25 Euro, approximately US$28) (Carrascal and de Oliveira 2013). Interestingly, the value of the two services together is roughly equal to that of the valuation of the data. This point means that it might be possible for a service provider to obtain the needed data to create her service for the cost of offering the public and the individual screening tests, with little or no monetary compensation to users.

There are several limitations in the way that the goal of data use was mentioned. First, we examined willingness to pay or be paid for a single service. In real-life applications, users might be offered screening for multiple diseases, which may increase the perceived value to participants. Additionally, we did not specify if our request for data was a one-time request or for ongoing access, nor the retrospective length of time that the data would be accessed. Future research will examine the effect of these variables on the valuation of the data and the offered services.

In many ways, the willingness to share one's personal data for personal and/or public use shares similarities with organ donation, in the sense that people might be more willing to share information with the possibility to help others if the default is sharing of such data. Past work has shown that the main difference among countries in the rates of organ donation can be explained not by kindness nor altruism, but in-

Table 2: Model coefficients for predicting monetary value. Two models are shown per condition: One where only the experimental conditions are used and the other where also user characteristics are included. Only non-censored observations are used.

Variable	User Pays				User is Paid			
	Coefficients (SE)	P-value	Coefficients (SE)	P-value	Coefficients (SE)	P-value	Coefficients (SE)	P-value
Severity	0.149 (0.171)	0.386	0.182 (0.168)	0.281	0.272 (0.102)	0.008	0.251 (0.102)	0.014
Personal	-0.063 (0.161)	0.696	-0.090 (0.160)	0.576	-0.274 (0.104)	0.009	-0.246 (0.104)	0.019
Public	-0.101 (0.165)	0.542	-0.131 (0.165)	0.430	-0.247 (0.102)	0.016	-0.236 (0.102)	0.021
Medical status			0.365 (0.133)	0.007			0.009 (0.081)	0.914
Education			0.066 (0.027)	0.017			0.004 (0.015)	0.811
Income			0.000 (0.000)	0.675			0.000 (0.000)	0.054
Age			-0.008 (0.006)	0.150			0.006 (0.003)	0.105
Gender			-0.020 (0.124)	0.870			0.033 (0.075)	0.659

stead in the consent form and default option. In many countries, implementing an Opt-Out Policy is the main reason for an increase in the rate of organ donations (Goldstein and Johnson 2003). Similarly, we hypothesize that if by default personal data will be used for uses such as the ones outlined in this work, and only if the user chooses to "opt-out" will it be removed; more often than not users will allow their information to be analyzed, potentially helping the greater good.

Moreover, the likelihood of one choosing to be a living liver donor increases significantly if a personal acquaintance is in need of a liver or was in need of it (Papachristou et al. 2004). The latter is similar to our results, indicating people who suffered from a medical condition themselves were willing to pay more for the use of their data than those who did not. However, care should be taken not to use data only from those who are biased towards data donation, as this could cause a bias in the models, both for the condition that they are suffering from and in other conditions.

We believe that when people can share their personal data for medical uses, one of the barriers that may prevent people from doing so would be the lack of supervision on the use of these data. This will be especially true if the default for the use of data will be Opt-In. Thus, we envision an independent organization, similar to Institutional Review Boards, which will oversee the proper use of data. Such an organization will weigh the use of personal information against the possibility of harm that can arise from its use, allowing personal information to be used (and possibly shared) when the greater cause outweighs the possible harm.

Acknowledgements

Omer Ben-Porat and Moshe Tennenholtz are supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 740435).

References

Allerhand, L.; Youngmann, B.; Yom-Tov, E.; and Arkadir, D. 2018. Detecting parkinson's disease from interactions with a search engine: Is expert knowledge sufficient? In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 1539–1542. ACM.

Ben-Sasson, A., and Yom-Tov, E. 2016. Online concerns of parents suspecting autism spectrum disorder in their child: content analysis of signs and automated prediction of risk. *Journal of medical Internet research* 18(11):e300.

Carrascal, Juan Pablo, R. C. E. V. C. M., and de Oliveira, R. 2013. Your browsing behavior for a big mac: Economics of personal information online. *Proceedings of the 22nd international conference on World Wide Web* 189–200.

Danezis, G.; Lewis, S.; and Anderson, R. 2005. How much is location privacy worth? *WEIS* 5.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

Debreu, G. 1959. Topological methods in cardinal utility theory. In *Mathematical methods in the social sciences*. Stanford University Press, Stanford. 16–26.

Dewar, K. 2017. The value exchange: Generating trust in the digital world. *Business Information Review* 34(2):96–100.

Dodge, Y. 2003. *The Oxford Dictionary of Statistical Terms*.

Goldstein, D., and Johnson, E. J. 2003. Do defaults save lives?

Hochberg, I.; Daoud, D.; Shehadeh, N.; and Yom-Tov, E. 2019. Can internet search engine queries be used to diagnose diabetes? analysis of archival search data. *Acta Diabetica*.

Kugler, L. 2018. The war over the value of personal data. *Communications of the ACM* 61(2):17–19.

Malgieri, G., and Custers, B. 2018. Pricing privacy–the right to know the value of your personal data. *Computer Law & Security Review* 34.2:289–303.

Misje, A. H.; Bosnes, V.; Gåsdal, O.; and Heier, H. E. 2005. Motivation, recruitment and retention of voluntary non-remunerated blood donors: A survey-based questionnaire study. *Vox sanguinis* 89.4:236–244.

Nisan, N., and Ronen, A. 2001. Algorithmic mechanism design. *Games and Economic behavior* 35(1-2):166–196.

Papachristou, C.; Walter, M.; Dietrich, K.; Danzer, G.; Klupp, J.; Klapp, B. F.; and Frommer, J. 2004. Motivation for living-donor liver transplantation from the donor's perspective: An in-depth qualitative research study. *Transplantation* 78.10:1506–1514.

Paparrizos, J.; White, R. W.; and Horvitz, E. 2016. Screening for pancreatic adenocarcinoma using signals from web search logs: Feasibility study and results. *Journal of Oncology Practice* 12(8):737–744.

Peer, E.; Brandimarte, L.; Samat, S.; and Acquisti, A. 2017. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70:153–163.

Skatova, A.; Ng, E.; and Goulding, J. 2014. Data donation: Sharing personal data for public good? *Application of Digital Innovation. London, England: N-Lab*.

Soldaini, L., and Yom-Tov, E. 2017. Inferring individual attributes from search engine queries and auxiliary information. In *Proceedings of the 26th international conference on World Wide Web*, 293–301. International World Wide Web Conferences Steering Committee.

Vickrey, W. 1961. Counterspeculations, auctions, and competitive sealed tenders. *Journal of Finance* 16:15–27.

White, R. W., and Horvitz, E. 2017. Evaluation of the feasibility of screening patients for early signs of lung carcinoma in web search logs. *JAMA Oncology* 3(3):398–401.

White, R. W.; Doraiswamy, P. M.; and Horvitz, E. 2018. Detecting neurodegenerative disorders from web search signals. *NPJ Digital Medicine* 1(1):8.

Wicks, P.; Vaughan, T. E.; Massagli, M. P.; and Heywood, J. 2011. Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nature Biotechnology* 29(5):411.

Yom-Tov, E. 2016. *Crowdsourced health: How what you do on the Internet will improve medicine*. MIT Press.

5 Full questionnaire and debrief statement

Questionnaire

Demographic Questions:

- What year were you born?
- What is your gender?
  - Male
  - Female
  - Other
  - I prefer not to disclose
- In which country do you currently reside?
- What is the highest level of school you have completed or the highest degree you have received?
  - Less than high school diploma
  - High school graduate
  - Some college but no degree
  - Associate degree in college (2 year)
  - Bachelor's degree in college (4 year)
  - Master's degree
  - Doctoral degree
  - Other
- What is your yearly income level?
  - Less than 15,000$
  - 15,000$-30,000$
  - 30,000$-45,000$
  - 45,000$-60,000$
  - 60,000$-75,000$
  - 75,000$-90,000$
  - 90,000$-115,000$
  - More than 115,000$
- Have you suffered from a serious medical condition in the past year?
  - Yes
  - No
- Are you currently suffering from a serious medical condition?
  - Yes
  - No

Experimental condition:

(Each participant received only one question from the following):

- We want to measure and report the rate of flu virus in your country.
- We want to develop a way to detect thyroid cancer by analyzing your Bing or Google search queries and apply it to people in your country.
- An algorithm will examine your data in order to test if it suggests that you have thyroid cancer and report the results to you.
- We want to develop a way to detect thyroid cancer using your Bing or Google search queries. We will examine your data in order to see if it suggests that you have thyroid cancer and report the results to you.

(Each participant received only one question from the following):

- To do so, we would like to purchase your Bing or Google search queries in order to analyze them. We plan to recruit 1000 people to complete this questionnaire. Since we only require 100 responses, we will only be contacting the 100 people who requested the least amount of money (as long as it is less than our maximum). For how much money (in US dollars) would you be willing to sell us your search queries for analysis?
- To do so, we would like to analyze your Bing or Google search queries. We plan to recruit 1000 people to complete this questionnaire. Since we only need 100 responses, we will contact the 100 people who were willing to pay the largest amount of money (as long as it is more than our minimum amount). How much money (in US dollars) are you willing to pay so that we analyze your search queries?

Debrief statement

Thank you for your participation in our study! Your participation is greatly appreciated.

As we noted in the description of our study, recent advances in medical research have shown that it is possible to diagnose serious medical conditions from the searches people make online through services such as Google and Bing. In the consent form of the study, we informed you that the

purpose of our study was to recruit a large group of people from whom we will collect their search history, which we will use to improve our ability to discover medical conditions.

In actuality, the goal of our study was to estimate the monetary value that users place on data, which can be used to assist in medical research. The way in which the questionnaire was structured is known to elicit truthful responses from people, but the minimal prices we intended to offer were set up so that no people could be recruited for the second stage of the research.

Unfortunately, to properly test our hypothesis, we could not reveal these details to you at the time of the experiment.

We hope that the results of our study will help us offer these novel screening services to people around the world in the near future. We thank you for helping us in understanding the value people ascribe to these services.