Risk, Trust, and Altruism in Genetic Data Sharing

Zeeshan Samad, Myrna Wooders, Bradley Malin, Yevgeniy Vorobeychik

August 14, 2020

Abstract

This study investigates individual attitudes toward privacy risks associated with the sharing of personal information. We conduct behavioral experiments to elicit attitudes with respect to several conditions. First, we consider two distinct scenarios to explore how types of information provided shape behavior. We examine two types of information: 1) genetic data shared with a healthcare provider; and 2) financial data shared with a money manager. In the former case, uncertain benefit is stated in terms of health outcomes, whereas in the latter, uncertain benefit is stated in terms of financial benefit. Both scenarios involve identical decisions and monetary stakes, permitting us to focus on how the framing of data sharing influences attitudes. Second, we design experiments to investigate the motivations behind decisions in terms of altruism and trust in data sharers. Third, we consider whether data recipients protect shared data when protection is costly and benefits data sharers only. Our findings (with 162 subjects) indicate that individuals are more willing to risk a loss to privacy of genetic data (for an anticipated return in health benefits) than they are to risk loss of financial data (for an anticipated return in financial benefits). We further observe that trust has a significant impact on the investment frame, but not on the genetic frame. Finally, we find that 50 to 60 percent of data recipients choose to protect another person's data, with no significant differences between frames.

1 Introduction

Cheap DNA sequencing has enabled broad collection, processing and sharing of genetic information [29]. This information is disclosed by patients to physicians, by research participants to scientists, and by general consumers to direct-to-genetic testing companies (e.g., 23andme.com) for various purposes. Genetic information sharing provides an individual with the opportunity to improve health and contribute to societal endeavors. At the same time, while some people share genetic information openly [19] (e.g., posting to websites such as OpenSNP [17] or GEDMatch [18]), many others consider such information to be highly personal and potentially sensitive [30]. As a result, many people prefer that their information be managed and used in a manner that preserves their expectation of privacy [14]. Sharing data with another party is not without risks. For instance, a data recipient may fail to protect a data sharer's privacy through weak anonymization practices or misuse the data due to lack of access control and institutional oversight.

Although various investigations have examined the extent to which people are concerned about the privacy and security of their genetic data, there are several major limitations to existing studies. First, they typically elicit information about privacy concerns through surveys [23, 24, 5, 12, 13, 27, 7, 28, 6], despite the fact that survey instruments often fall prey to the *privacy paradox*, whereby reported attitudes about privacy end up being inconsistent with measures people actually take to protect their data [2]. Second, existing research often neglects the fact that data sharing requires an interaction between two parties (e.g., a patient and a physician), which can influence behavior.

A natural way to address these limitations is to adopt an economic perspective, whereby a genetic data sharer's decision is modeled as a risky investment. In this representation, sharing genetic data may result in long-term benefit (e.g., treatment for an undiagnosed condition [3]), but may expose an individual to privacy risks. Similarly, a data recipient's decision about how and when to protect genetic data has a salient economic aspect, as data protection can be costly and data recipients often have a limited budget. In both cases, a decision has ramifications for the other party.

In this work, we investigate whether decisions about sharing genetic data are different from decisions about sharing financial data, and if this information can be used to better understand how individuals value privacy of their genetic data relative to the privacy of their financial data. Better understanding is crucial to determining the conditions under which people share genetic data, and how such sharing can be encouraged.

We conduct this investigation of the implications of economic decisions involved in genetic data sharing and protection through controlled experiments with human subjects, focusing particularly on the impact of *framing*, which has been shown to affect decisions in a number of other behavioral studies [15, 25, 22]. Specifically, we represent the problem in two ways. First, we design a scenario where a *patient* can undergo a *genetic test*, and thereby share *genetic* data with a *physician*. Second, we design a scenario in which an *investor* can *invest in a risky asset*, thereby sharing *financial* data with a *money manager* purely for financial gain. We refer to these scenarios as the *genetic frame* and the *investment frame*, respectively. Crucially, actual monetary gain and loss potentials are identical in both frames.

We expand on this experimental design through several variations that allow us to isolate the effects on decisions of economic and altruistic motivations, as well as trust and reciprocity. To do so, at the beginning of each experiment, subjects, acting as patients or investors, are provided with a fixed amount of money as an endowment. Each subject makes a decision on whether to share data, that is, whether or not to undergo a genetic test (in the genetic frame) or whether or not to make an investment (in the investment frame). If a subject does not share data, she keeps her endowment as her final payout. If she does share data, the subject faces two uncertain possibilities, one positive and the other negative. First, with a small probability, she may receive a substantial sum of money (i.e., a return on investment), representing treatment in the genetic frame or financial gain in the investment frame. Second, with a larger probability, the subject loses *all* of her endowment, representing the risk associated with a potential privacy compromise of shared data as a result of her decision.

The design of the aforementioned experimental setting effectively removes the role of the data recipient from the equation, thus casting the problem purely as a risky decision. While the presence of the data recipient does not change the economic nature of this choice, one may expect that people will not act out of pure self-interest. If data sharing can benefit another party, it is possible that this will serve as an incentive to share. This motivates an *altruism setting*, in which data sharing also benefits a data recipient.

Finally, we establish a *trust setting* in which data sharing benefits both the data sharer and the data recipient but, in addition, the data recipient can, at a cost, reduce the data sharer's risk of privacy loss. This enables us to measure the extent to which trust in the recipient's reciprocity motivates a greater degree of data sharing. Thus, all data sharers (i.e., patients or investors) decide whether or not to share data in three different settings, namely base, altruism, and trust.

Turning our attention to data recipients, we also study their motivations to reduce the privacy risk faced by data sharers. Data recipients decide whether or not to bear a monetary cost that would reduce the data sharer's risk. They make this decision in two settings, a *reciprocity setting* and a *recipient-altruism* setting. In the reciprocity setting, the data sharer's decision to share data has a direct financial benefit on the recipient, independent of the latter's decision to reduce data privacy exposure risk. In the recipient-altruism setting, the data recipient starts with data and can spend money to reduce the risk of loss to the data sharer, to whom the data belongs according to the experiment narrative. Comparing these two settings allows us to isolate the effects of altruism and reciprocity on motivating the data recipient to reduce the sharer's risk.

Our findings show that data sharers are significantly more likely to tolerate risk when sharing genetic data than when sharing financial data. We further find that trust and reciprocity appear to encourage sharing more than altruism does.

2 Methods

We study data sharing through human subject experiments involving monetary stakes. Our experiments involve two classes of settings: a decision under uncertainty involving a single subject (e.g., the base setting), and an interaction between two subjects (e.g., the trust & reciprocity setting). In the latter class of settings, we first divide subjects into data sharer and data recipient pairs, randomly and anonymously. Data sharers can share their data and data recipients can expend resources to protect that data and reduce data sharers' risks.

Each {data sharer, data recipient} pair is randomly assigned to either the genetic frame or the investment frame. In the genetic frame, data sharers are assigned to a role of a patient and data recipients are assigned to the role of a physician. In the investment frame, data sharers are assigned to a role of an investor and data recipients are assigned to a role of a money manager. Thus, we pair patients with physicians in the genetic frame, and investors with money managers in the investment frame. All subjects maintain their randomly assigned roles in all experimental settings.

The underlying decision problem - in terms of monetary payoffs, risks, and choices - is the same in both frames. The only difference is what the numbers represent. In the genetic frame, the decision problem simulates a situation where a patient needs to decide whether or not to undergo a genetic test. In the investment frame, the decision problem simulates a situation where an investor needs to decide whether or not to make a risky investment. The risk in either case stems from potential privacy breaches associated with data sharing by the patient or the investor.

Within each frame, we conduct four versions, i.e., settings, of the experiment that allow us to deconstruct

Table 1: Summary of differences among settings

Setting name	Sharers' motivation	Recipients' motivation
Base	1. Personal benefit	N/A
Altruism	1. Personal benefit 2. Altruism	N/A
Trust & Reciprocity	1. Personal benefit 2. Altruism 3. Trust	1. Altruism 2. Reciprocity
Recipient- Altruism	1. Personal benefit 2. Altruism 3. Trust	1. Altruism

In the base setting, data recipients do not even exist. In the *altruism* setting, data recipients participate only as passive recipients of their data sharer's altruism. Thus, data recipients do not have any actions in the *base* and *altruism* settings. The *trust and reciprocity* setting is the full version of the game with the highest number of confounding motivations. In the *recipient-altruism* setting, a data recipient is given money from the experimenter, regardless of the data sharer's action. Note that in both the *trust and reciprocity* setting and the *recipient-altruism* setting, a data recipient has an action only if the data sharer chooses to share data.

the various motivations that individuals may have to share their data, similar to a triadic design approach pioneered by Cox [8, 10, 9, 11] for decomposing motivations. We hypothesize that data sharers have three reasons to share their data: (i) potential health/financial benefits resulting from genetic testing/investment; (ii) benefits to data recipients, such as money managers or healthcare institutions/physicians who are engaged in research; and (iii) trust in the ultimate data custodian to protect their data from being stolen or misused. The four settings of our experiment are called : 1) base, 2) altruism, 3) trust and reciprocity, and 4) recipient-altruism. Data sharers participate in the first three settings, i.e., 1) base, 2) altruism, and 3) trust and reciprocity. Data recipients participate in the last three settings, i.e., 2) altruism, 3) trust and reciprocity, and 4) recipient-altruism. Below, we describe each setting in detail, while Table 2 provides a summary of the settings and the associated motivations that are tested.

2.1 Base setting

In this setting, a data sharer is endowed with \$6 and must choose whether or not to share her data, which costs \$2 out of the \$6 endowment. If she chooses to share data, she can win \$60 with a 5% probability, but also lose her remaining \$4 with a 25% probability. Note that there is no mention of a data recipient in this setting, so other-regarding preferences cannot affect a data sharer's decision.

Patients are told that the \$60 represents potential health benefits from genetic testing and the loss of \$4 represents the loss of privacy of genetic data. Investors are told that \$60 represents a financial gain from the investment while \$4 represents loss of privacy of financial data. If a person's genetic data gets hacked and/or misused, she could potentially lose her health insurance, and in the worst case, her job and family ties, depending on the information revealed through the genetic test. To capture this extreme case, we represent the cost of exposure of sensitive data as the loss of all wealth. In order to obtain meaningful results, we deliberately give participants a higher risk of their data being exposed than what it is in the real world. The fact that we do not use real-world numbers does not create a problem because our goal is not to estimate the proportion of population that is willing to perform a genetic test given awareness of actual risks. Rather, our objective is to compare the impact of privacy concerns when they pertain to genetic versus financial data. Financial data, in turn, are natural baselines for comparison, as this is the most common frame used in behavioral economic studies.

2.2 Altruism setting

Typically genetic testing and investment settings are not single-agent decisions, but involve another party such as physicians or money managers. In such encounters, a number of factors may influence people's decisions in addition to egocentric motivations, including altruism, trust, and reciprocity. In order to tease apart these factors, we introduce an *altruism setting.* In this setting, a data sharer's decision problem is the same as in the base setting with one exception. Specifically, if a data sharer chooses to share data and give up \$2, a data recipient (who is said to be a physician or a money manager, depending on the frame) receives \$4. This \$4 represents either the physician's benefit from advancing his research or the money manager's benefit. Note that while a data recipient receives money, he does not have a decision to make. This setting includes two motivations for a data sharer to share data: (i) her potential benefit (which is also present in the base setting) and (ii) her desire to provide benefit to a data recipient, or altruism. We hypothesize that if altruism is a motivation for sharing data, then a data sharer will be more likely to share data in the altruism setting than in the base setting.

2.3 Trust and Reciprocity setting

For data sharers, this setting is called the trust setting and for data recipients, it is called the reciprocity setting. In the experiment, these separate settings are implemented as a single, interactive game in which both parties make decisions that affect their own as well as each other's payoffs. This setting resembles the extensively-studied trust game [4] in which a player, say Player 1, is endowed with some money and chooses whether to send a portion of it to another player, say Player 2. Player 2 receives three times the amount sent by Player 1, and then has the opportunity to send some of his money back to Player 1. For example, if Player 1 sends \$5 from her \$10 endowment to Player 2, then Player 2 receives \$20. Player 2 can then send some of his \$20 back to Player 1. There are robust findings in this game: (a) Player 1 often sends a non-zero amount to Player 2, and (b) Player 2 commonly reciprocates and sends a portion of her funds back to Player 1.

The trust and reciprocity setting adds one more feature to the altruism setting. Specifically, a data recipient may now spend some of his \$4 to reduce the data sharer's risk of loss. Specifically, a data recipient has three options: spend nothing (i.e., \$0), spend half of his earnings (i.e., \$2), or spend all of his earnings (i.e., \$4). If a data recipient spends \$0, then the data sharer faces a 25% probability of losing her remaining wealth of \$4 (which is the same probability as in the other settings); if a data recipient spends \$2, then the data sharer faces a 15% probability of losing her \$4 wealth; and lastly, if a data recipient spends all of his \$4, the data sharer faces no risk (i.e., a 0% probability) of losing her \$4 wealth.

In the base and altruism settings, a data sharer's expected payoff is \$6 from sharing data and \$6 from not sharing data (thus, a risk neutral person would be ambivalent between the two choices). Now a data sharer's expected payoff depends on a data recipient's decision. If a data sharer believes that the data recipient will spend his entire \$4, then her expected payoff is \$7. If she believes that a data recipient will spend \$2, then her expected payoff is \$6.40. Finally, if she does not expect the data recipient to spend anything, then her expected payoff is \$6, the same as in the other settings. Therefore, a data sharer now has three motivations to share data: (i) her own benefit (this motivation is present in all three settings), (ii) her altruism, or desire to benefit a data recipient (this is also present in the altruism setting but is not present in the base setting), and (iii) whether she trusts her data recipient to protect her wealth (which is not present in either of the previous two settings). If a data sharer is more likely to share data in the trust setting than in the altruism setting, then trust is what accounts for the difference.

Let us now turn shift our focus to the motivations of a data recipient. If a data recipient spends a positive amount to reduce a data sharer's probability of loss, then there are two possible factors that could be motivating this choice: (i) altruism: he wants to protect the data sharer from a loss; and (ii) reciprocity: he feels beholden from knowing that his income is a result of the data sharer's decision, and thus reciprocates the favor by spending some money to help the data sharer. To disentangle these two factors, we construct a *recipient-altruism* setting that is described below.

2.4 Recipient-altruism setting

The purpose of this setting is to deconstruct the different underlying motivations of data recipients (i.e., physicians and money managers) to spend money to protect a data sharer's data and, consequently, her endowment. The only difference between this recipient-altruism setting and the reciprocity setting is that the data recipient's earning of $4 is not a consequence of the data sharer's decision to share data. Instead, data recipients are now endowed with $4 regardless of data sharers' choice. Each data recipient is told that a data sharer's personal data, and hence wealth, is at risk and data recipients need to decide whether to spend some of their $4 to protect it. Thus, a data recipient only has one motivation to spend money to protect a data sharer's wealth: altruism. The increase in the tendency of data recipients to protect data sharers' data in the reciprocity setting compared to this altruism setting would then be motivated by reciprocity. Note that in both reciprocity and recipient-altruism settings, data recipients have no purely economic motivation to spend money to protect a data sharer's endowment.

2.5 Implementation

We conducted several experiment sessions at Vanderbilt University in March 2018 and October 2018, recruiting a total of 162 undergraduate students to participate in the experiment. All sessions took place on weekdays in a computer lab that had 30 computer stations, and subjects submitted all responses using computers. Subjects were paid in cash for all settings at the end of the experiment session. While payments were being calculated and put into envelopes, subjects were afforded the opportunity to complete a demographic questionnaire.

We conducted a total of 8 sessions, 4 for the genetic frame and 4 for the investment frame, with 78 subjects in the genetic frame (39 patient and physician pairs) and 84 subjects in the investment frame (42 investor and money manager pairs). Each data sharer participated in the first three settings, i.e., 1) base, 2) altruism, and 3) trust and reciprocity, while each data recipient participated in the last three settings, i.e., 2) altruism, 3) trust and reciprocity, and 4) recipient-altruism. Subjects were not informed about their outcomes, and therefore also earnings, from each setting until the end of the experiment session.

At the beginning of the experiment, subjects were informed that they would be participating in multiple decision-making tasks, but were not informed about the exact number or nature of the tasks. They only learned about each task at the beginning of each setting. This helped us ensure that each subject's first decision remained pure from any undesirable behavioral effects, such as priming or portfolio effects.

To ensure that responses were free from such effects, we gave these settings to subjects in various sequences (i.e., some subjects were given the base setting first, some the altruism setting, and others the trust setting). Any undesirable effects would have contaminated the results in the second (or later) setting in which a subject participated. In the event of such contamination, we would have discarded subjects' responses in the second and third settings, keeping only the responses recorded from the first setting, which would be guaranteed to be free from such effects.

We found that the order in which a subject received a setting did not make a difference in the proportion of subjects who shared their data. For example, the proportion of patients who chose to get a genetic test in, say, the base setting remained the same regardless of whether they participated in the base setting as their first, second, or third setting. Therefore, we conclude that subjects made their decision in each setting independently of their decision in another setting. This implies that our results are free from any undesirable effects (e.g., priming or portfolio effects) and we can keep responses from all settings.

To ensure anonymity while paying everyone their correct earnings, we provided each experiment participant a unique numerical code, which was used as their pseudonym throughout the experiment. All subjects were guaranteed minimum earnings of $5 for participating, but could earn substantially more, depending on both their choices and the choices of the participants with whom they were paired. Average subject earnings were approximately $20, and average time spent was approximately 45 minutes (including administrative tasks such as introduction, questions, and payments).

Table 2 provides summary statistics for the demographic characteristics of subjects. The fact that none of the subjects' characteristics are different across the two frames indicates that assignment to frames was sufficiently random. . There is no statistically significant difference between the two groups in terms of age, gender, race, religiosity, or being a victim of identity theft. Only the variable real-life investment is statistically different across the two groups, but this is because this question is actually different for each group. For the genetic group, the question asks subjects whether they have undergone genetic testing in real life, whereas for the investment group, the question asks whether they have made any investments in real life.

Table 2: Summary Statistics

Variables	Investment	Genetic	Difference
Age	19.71	20.03	-0.326
Victim of ID theft	0.076	0.029	0.033
Real-life Choice	0.519	0.116	0.409***
% Female	0.506	0.406	0.133
% White	0.544	0.580	-0.003
% Hispanic	0.101	0.130	-0.030
Religiousness (0-3)	1.329	1.304	-0.060

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

This table reports means of demographic characteristics, separately for the genetic and investment groups. The last column shows the difference in means. A statistically significant difference indicates that subjects in the genetic and investment groups are different in terms of that characteristic. The variable *Real-life Choice* is different across the two groups, but that is not a concern because this variable represents an entirely different question for each group. For the investment group, it represents whether subjects have made personal investments in real life while for the genetic frame it represents whether subjects have ever undergone genetic testing in real life. Thus, overall the assignment into groups was perfectly randomized. The p-values reported in this table are based on two-tailed t-test results (we also conducted χ^2 and Mann-Whitney tests, both of which provide the same conclusions).

2.6 Online Replication

We also replicate the laboratory experiment using the online platform Amazon Mechanical Turk. The results of the online experiment are highly similar to the in-person laboratory experiments, but we chose not to pool data from the two experiments because the subject populations and experimental environments are different. Rather, the online settings demonstrate robustness of our results to different settings and subject populations. We refer the reader to the Supplementary Information for details of this experiment.

3 Results

We first study the decision of an individual with respect to whether or not to share data with another party. In the genetic frame, a subject who is assigned the role of a patient needs to decide whether to take a genetic test and share her genetic data with a physician. In the investment frame, a subject assigned the role of an investor needs to decide whether to invest in an asset and share financial data with a money manager. In both frames, i.e., regardless of the narrative presented to the subject, experiment outcomes are monetary, determined as follows. Subjects are endowed with $6 at the beginning of an experiment. If they choose not to share data (either by getting a genetic test or by making an investment, depending on the frame), they receive the full endowment at the end of the experiment. If instead they choose to share their data, they give up $2 from their endowment to face an uncertain outcome that is defined as follows. On the one hand, data sharers have a 5% probability of receiving \$60 (representing either a health or financial benefit, depending on the frame), for a total final payout of \$64 at the end of the experiment. On the other hand, they have a 25% probability of losing their remaining endowment, representing the consequence of the loss of privacy of shared data, leaving them with nothing at the end of the experiment. Note that from a purely decision-theoretic viewpoint, the two decisions have identical expected earnings of \$6; consequently, a risk-neutral individual would be indifferent between either sharing her data or not sharing her data, whereas a risk averse person would not share data.

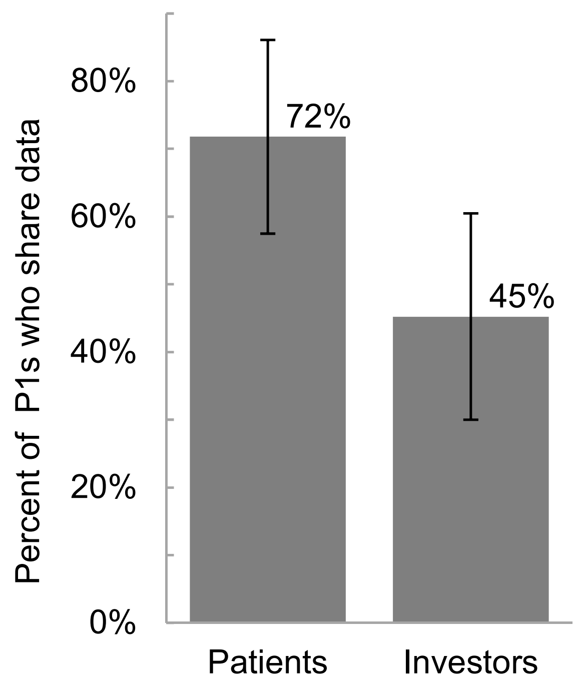
More People Choose to Share Data in the Genetic Frame In the base setting, we remove the data recipient from the equation. Instead, this setting involves only a data sharer who decides whether or not to put her data (genetic or financial) at risk, without explicit reference to a data recipient. Our main result in the base setting, shown in Figure 3, is that there is a substantial framing effect. Specifically, 72% of subjects in the genetic frame choose to share data, strongly suggesting that the uncertain outcome is seen as favorable. By contrast, only 45% of subjects in the investment frame choose to share data, suggesting that the uncertain outcome is seen as somewhat unfavorable.

In addition to comparing the means using a t-test, we compare the two frames through a regression analysis, where we control for various demographic characteristics such as age, race, gender, degree of religious beliefs, and whether subjects have ever been victims of data theft. Table 3 presents these regression results. The first column of the table shows results produced by an OLS regression and the second column shows results produced by a Probit regression. In both regressions, the dependent variable is the percentage of subjects who choose to share data (framed as choosing either to undergo a genetic test or to make an investment).

The key independent variable is *Genetic Frame*, which is a binary variable indicating whether a subject is given the genetic or investment frame. In the OLS regression, the variable Genetic Frame has a coefficient of 0.254, which suggests that subjects in the genetic frame (i.e., patients) are about 25.4 percentage points more likely to get a genetic test than investors are likely to make an investment. In the Probit regression, a marginal analysis shows that patients choose to undergo a genetic test with a probability of 0.686 while investors choose to make an investment with a probability of 0.425, which is a difference of 26.1 percentage points.[1] Thus, the results of Figure 3 are broadly consistent with the OLS and Probit results. The regression

[1]The probit coefficient shown in the table, 0.675, is the difference between the z-scores of the genetic frame and of the investment frame. We use this coefficient to perform a marginal analysis whose interpretation is more meaningful for our purposes.

Figure 1: Framing Effects



This bar chart shows framing effects in the base setting (in which patients/investors are asked to make decisions without any mention of physicians/money managers). We find that 72% of subjects in the genetic frame share their data, compared to only 45% of subjects in the investment frame. This difference is statistically significant ($p = 0.015$ for a two-tailed t-test). Error bars represent 95% confidence interval.

Table 3: Regression results

Variables	OLS	Probit
Genetic Frame	0.254**	0.675**
Failed control questions	0.0243	0.0768
Age	-0.0389	-0.111
Victim of ID theft	0.174	0.493
% Female	0.0522	0.123
% White	0.169	0.466
% Hispanic	-0.216	-0.584
Religiousness (0-3)	-0.0101	-0.0302
Constant	1.089	1.716

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

This table shows results of OLS and Probit regressions. In both regressions, the dependent variable is the percentage of subjects who share data. The variable *Genetic Frame* is statistically significant. This is a binary variable indicating whether a subject is assigned to the genetic frame or investment frame. The results show that patients (i.e., subjects in the genetic frame) are about 25 percentage points more likely to get a genetic test than investors are likely to make an investment. Other variables do not significantly affect the decision to share data.

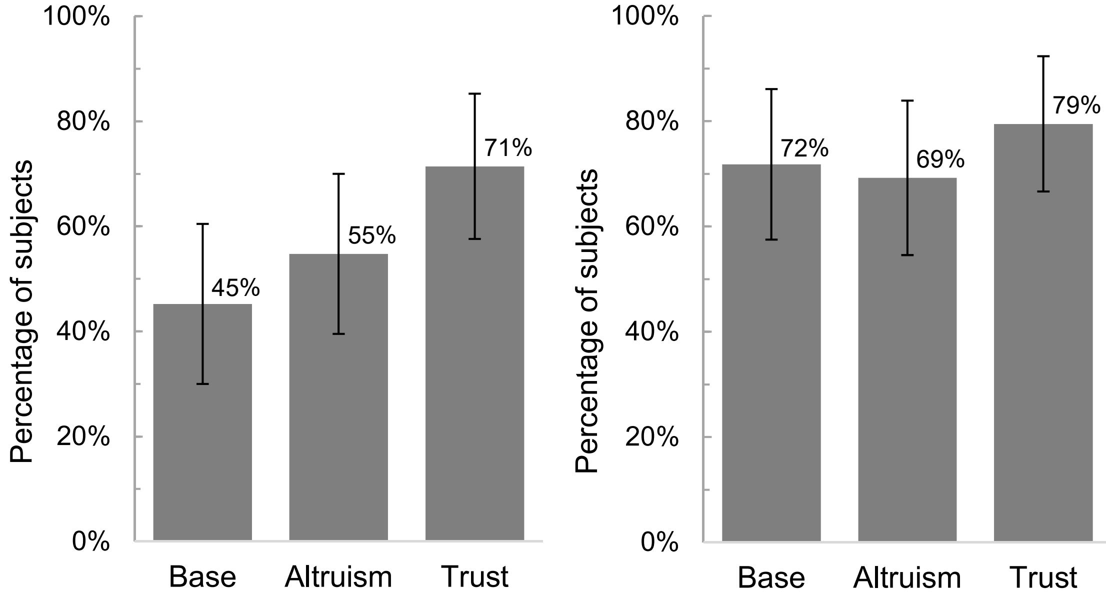
results also show that none of the demographic variables has a significant influence on the decision to share data, suggesting that people's willingness to put data (whether genetic or financial) at risk is similar across different age groups, gender groups, and racial groups.

Trust Increases Data Sharing Next, we investigate whether the involvement of a data recipient influences the decision to get a genetic test or make an investment. The presence of a data recipient/custodian (who may be an individual or an organization) introduces considerations into the decision about whether data is shared. These may include other-regarding preferences as well as a degree of trust in the data custodian's willingness to protect data. Thus, the following two settings consider the impact of these considerations, in terms of *altruism* and *trust*, on a data sharer's decision as to share data.

In the altruism setting, data sharers (i.e., patients and investors) are randomly and anonymously paired with data recipients (i.e., physicians and money managers). By choosing to share data, the data sharer *also benefits the data recipient*, who receives \$4 if the data sharer decides to share data. This gain represents investment gain to the money manager in the investment frame and helps the physician with biomedical research in the genetic frame.

The trust setting also features a data recipient (i.e., a physician or a money manager) who receives \$4 if the data recipient shares data. However, now the recipient is viewed as a data custodian who may choose to protect the sharer's data. In particular, the data recipient now has three options: (i) spend \$0 to protect the data, which keeps the data sharer's risk of loss at 25%; (ii) spend \$2 to protect the data, which reduces

Figure 2: Choices made by Data Sharers in each setting



This chart shows the proportion of subjects who share data in each setting and each frame. In the genetic frame (left panel), there is no statistically significant difference between the heights of any two bars. In the investment frame (right panel), two pairwise differences are statistically significant: (i) between the trust and altruism settings (p=0.018) and (ii) between the trust and base settings (p=0.003). That is, when money managers are able to protect investors' data, investors are much more likely to invest. This suggests that investors trust money managers to protect their financial data and that trust is an important factor in their decision to invest. By contrast, when physicians are able to protect patients' data, it does not increase patients' likelihood to get a genetic test. This suggests that trust is not an important factor for patients in their decisions to undergo genetic testing. Rather, personal benefit is the only important motivator for getting a genetic test.

the data sharer's risk of loss to 15%; and (iii) spend all $4, reducing the sharer's risk to 0%. Thus, if the data sharer trusts the recipient to spend at least some of the money received for better data protection (concretely, reducing the risk of loss to the sharer), the data sharing option becomes far more appealing.

Figure 3 compares the results of all three settings in which a data sharer chooses to share data. While there is no statistically significant difference between any pair of settings in the genetic frame, there are statistically significant differences between settings in the investment frame. Specifically, in the trust setting, a significantly greater number of investors choose to share data, compared to those in the altruism and base settings. This implies that investors are more likely to invest when their money manager has the ability to reduce their risk, indicating that investors trust their money managers to have their best interests at heart.

Nevertheless, altruism increases the frequency of data sharing in the investment frame by 10 percentage points (from 45% to 55%). This is in stark contrast to the genetic frame, where altruism decreases the frequency of data sharing by 3 percentage points (from 72% to 69%). Indeed, our results in the genetic frame contrast with typical results in related *trust games* [4], where a robust finding is that altruism accounts for a large portion of investment [8]. Our results suggest a potential displacement effect [16] because fewer subjects share data in the altruism setting than in the base setting, even though their purely economic motivations are identical.

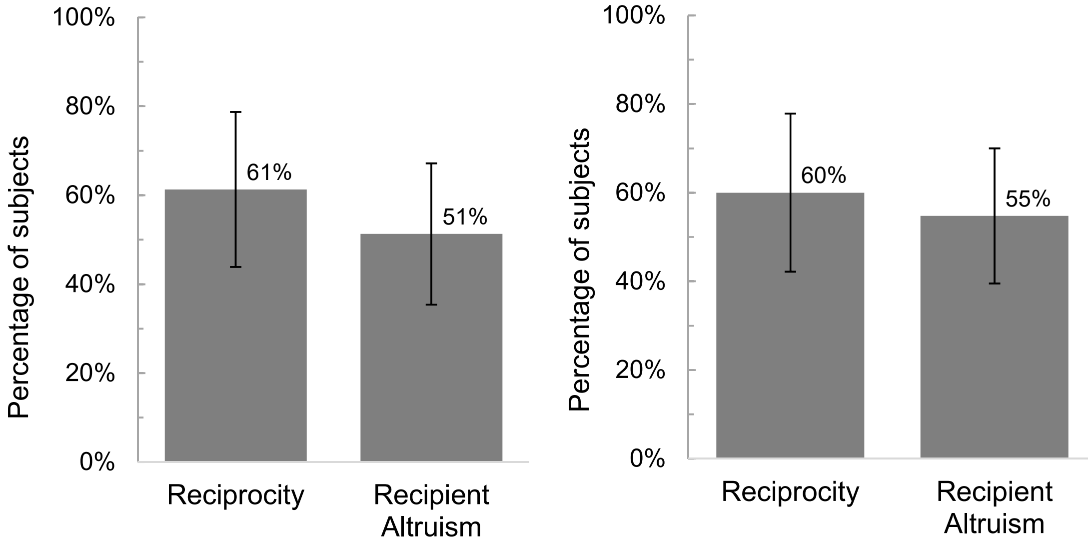
Approximately 50-60% of Data Recipients Choose to Protect Data One of the unique features of our study design is the interaction between a patient and a physician in the genetic frame and an investor and a money manager in the investment frame. Data recipients (physicians and money managers) may spend some of the money they receive to reduce the data sharer's risk of loss. This is modeled in our experiment as a reduction of risk of loss for the data sharer from 25% to either 15% or 0%. Based solely on economic self-interest, data recipients would not spend any of the money they receive on data protection. Nevertheless, they do.

Two additional factors, however, can motivate data recipients to protect the sharer's data: altruism and reciprocity. As these two motivations for a non-zero contribution by the data custodian cannot be directly disentangled in the trust setting, we introduce a *recipient-altruism* setting. In this setting, the data recipient receives $4 from the experimenter rather than the data sharer, and may still use it to reduce the data sharer's risk of loss. Thus, any money spent by the data recipient in the altruism setting is motivated solely by altruism.

The previously termed trust setting then becomes the *reciprocity* setting when viewed from the perspective of the data recipient, as now any *additional* spending beyond what we observe in the altruism setting must be due to reciprocity. The distinction between altruism and reciprocity is significant in practice as well: data custodians are often different from people who directly interact with the individuals sharing data (particularly as individuals in charge of data protection may change over time), and the nature of their decisions is best captured by the recipient-altruism setting.

Since spending the full $4 to protect the sharer's data is exceedingly rare, we pool all non-zero contributions by the data recipient into a single *decision to protect*, with the complement corresponding to spending nothing for the sharer's protection. Figure 3 presents the fraction of physicians and money managers who decide to spend some of their received funds on data protection, which benefits the data sharer. It can be seen that differences between the two frames are not dramatic, although the stronger result is encountered in the genetic frame. In the genetic frame, only 51% of data recipients choose to spend *anything* into reducing the risk to the data sharer when the sole motivation is altruism. In the investment frame, 55% of data recipients choose to spend anything to reduce the sharer's risk. The difference between these two

Figure 3: Choices made by Data Recipients in each setting



Recall that a data recipient assumes the role of a physician in the genetic frame and the role of a money manager in the investment frame. When reciprocity is a motivator, it increases the proportion of physicians who spend some money to protect their patients' data by 10 percentage points (61% - 51%) and the proportion of money managers who spend money to protect their clients' financial data by 5 percentage points (60% - 55%). In either frame, the difference is not statistically significant in a two-tailed comparison. However, a one-tailed t-test results in a marginally statistically significant difference in the genetic frame ($p = 0.05$). This suggests that reciprocity motivates physicians to spend some money to protect their patients' genetic data, but it does not motivate money managers to spend money to protect their clients' financial data.

numbers is not statistically significant. Reciprocity boosts the data recipient's spending from 51% to 61% in the genetic frame, and from 55% to 60% in the investment frame. The former comparison is marginally significant (one-tailed test; $p = 0.0516$). Nevertheless, even reciprocity is insufficient motivation because over one-third of data recipients still do not spend money to protect the sharer's data.

4 Discussion

There has been much public discussion about tension between data sharing and privacy in general [1], as well as about how privacy affects the sharing of genetic data in particular [28]. On the one hand, the ability to share fine-grained data is crucial to both research and policy [26]. On the other, privacy concerns must necessarily put a check on what, how, and with whom data is shared [21]. An important aspect of this discussion is the perceived risk to privacy of shared data, and how perceptions of risk affect trust. Perceived degrees of risk influence individual decisions to opt in or opt out of participating in clinical studies or research programs that involve shared datasets [20].

In this paper, we use methods from behavioral economics to understand how people make decisions about sharing data as well as the extent to which data custodians are motivated to protect it. Two especially important factors we explore are: 1) the effect of framing a decision in terms of risky investment; and 2) the effect of viewing data sharing explicitly as an encounter between two parties, a data sharer and a data recipient/custodian.

Our investigation yields several notable findings. First, despite identical monetary stakes, far more people choose to share their data when the decision is narrated as involving genetic testing, compared to when it is cast as financial in nature. As the distinction is solely one of framing, our finding suggests that people are more willing to take on risk when sharing genetic data than when sharing financial data. Second, it is noteworthy that both altruism and trust are significantly stronger motivators for sharing financial data than for sharing genetic data. Third, our finding that 50-60% of data custodians bear a cost to reduce a sharer's risk is, on the one hand, remarkable considering that such decisions only benefit data sharers and not the custodians themselves. On the other hand, social considerations suggest that without additional incentives or explicit regulation, shared data may be inadequately protected.

There are several limitations to our investigation that we wish to highlight, which stem from the fact that our experiments were naturally highly stylized, and thus lack several aspects of actual data sharing settings. The first of these is that the economic nature of the experiment means that ultimately the design revolves around monetary payouts, whatever the narrative. In reality, differences between genetic and financial data sharing situations also involve differences in stakes and in the nature of outcomes. Second, data sharing is often an encounter between an individual and an organization (e.g., individuals may opt into or out of data sharing) or among organizations, whereas our experiment focuses either on an individual's decision or on an encounter between two individuals. Third, data sharing encounters are typically not anonymous, whereas our experiment pairs subjects anonymously. Fourth, data sharing is largely implicitly captured through associated risks, whereas the actual decision is presented as making a financial investment or undergoing genetic testing. Despite these limitations, our observations provide insights into the nature of data sharing encounters and into the variations in behavior that, depending on the nature of data being shared, we should expect.

References

[1] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015.

[2] Susanne Barth and Menno DT De Jong. The privacy paradox - investigating discrepancies between expressed privacy concerns and actual online behavior - a systematic literature review. *Telematics and Informatics*, 34(7):1038–1058, 2017.

[3] Lisa Bastarache, Jacob J Hughey, Scott Hebbring, Joy Marlo, Wanke Zhao, Wanting T Ho, Sara L Van Driest, Tracy L McGregor, Jonathan D Mosley, Quinn S Wells, et al. Phenotype risk scores identify patients with unrecognized mendelian disease patterns. *Science*, 359(6381):1233–1239, 2018.

[4] Joyce Berg, John Dickhaut, and Kevin McCabe. Trust, reciprocity, and social history. *Games and economic behavior*, 10(1):122–142, 1995.

[5] Jeffrey R Botkin, Erin Rothwell, Rebecca Anderson, Louisa Stark, Aaron Goldenberg, Michelle Lewis, Matthew Burbank, and Bob Wong. Public attitudes regarding the use of residual newborn screening specimens for research. *Pediatrics*, 129(2):231, 2012.

[6] Ellen W Clayton, Colin M Halverson, Nila A Sathe, and Bradley A Malin. A systematic literature review of individuals' perspectives on privacy and genetic information in the united states. *PloS one*, 13(10):e0204417, 2018.

[7] Celeste M Condit, Lijiang Shen, Karen L Edwards, Deborah J Bowen, Diane M Korngiebel, and Catherine O Johnson. Participants' role expectations in genetics research and re-consent: Revising the theory and methods of mental models research relating to roles. *Journal of health communication*, 21(sup2):16–24, 2016.

[8] James C Cox. How to identify trust and reciprocity. *Games and economic behavior*, 46(2):260–281, 2004.

[9] James C Cox, Klarita Sadiraj, and Vjollca Sadiraj. Implications of trust, fear, and reciprocity for modeling economic behavior. *Experimental Economics*, 11(1):1–24, 2008.

[10] James C Cox, Vjollca Sadiraj, and Ulrich Schmidt. Paradoxes and mechanisms for choice under risk. *Experimental Economics*, 18(2):215–250, 2015.

[11] Giovanni Di Bartolomeo and Stefano Papa. Trust and reciprocity: extensions and robustness of triadic design. *Experimental Economics*, 19(1):100–115, 2016.

[12] D Duquette, C Langbo, J Bach, and M Kleyn. Michigan biotrust for health: public support for using residual dried blood spot samples for health research. *Public Health Genomics*, 15(3-4):146–155, 2012.

[13] Karen L Edwards, Amy A Lemke, Susan B Trinidad, Susan M Lewis, Helene Starks, Katherine W Snapinn, Mary Quinn Griffin, Georgia L Wiesner, and Wylie Burke. Genetics researchers' and irb professionals' attitudes toward genetic research review: a comparative analysis. *Genetics in medicine*, 14(2):236, 2012.

[14] Yaniv Erlich and Arvind Narayanan. Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics*, 15(6):409–421, 2014.

[15] Nancy S Fagley and Paul M Miller. Framing effects and arenas of choice: Your money or your life? *Organizational behavior and human decision processes*, 71(3):355–373, 1997.

[16] Uri Gneezy and Aldo Rustichini. A fine is a price. *The Journal of Legal Studies*, 29(1):1–17, 2000.

[17] Bastian Greshake, Philipp E Bayer, Helge Rausch, and Julia Reda. Opensnp–a crowdsourced web resource for personal genomics. *PLoS One*, 9(3), 2014.

[18] Ellen Greytak, David H Kaye, Bruce Budowle, CeCe Moore, and Steven Armentrout. Privacy and genetic genealogy data. *Science*, 361:857, 2018.

[19] Tobias Haeusermann, Bastian Greshake, Alessandro Blasimme, Darja Irdam, Martin Richards, and Effy Vayena. Open sharing of genomic data: Who does it and why? *PLoS One*, 12(5), 2017.

[20] Sara Chandros Hull, Richard R Sharp, Jeffrey R Botkin, Mark Brown, Mark Hughes, Jeremy Sugarman, Debra Schwinn, Pamela Sankar, Dragana Bolcic-Jankovic, Brian R Clarridge, et al. Patients' views on identifiability of samples and informed consent for genetic research. *The American Journal of Bioethics*, 8(10):62–70, 2008.

[21] Yann Joly, Edward S Dove, Bartha M Knoppers, Martin Bobrow, and Don Chalmers. Data sharing in the post-genomic world: the experience of the international cancer genome consortium (icgc) data access compliance office (daco). *PLoS computational biology*, 8(7), 2012.

[22] Jerwen Jou, James Shanteau, and Richard Jackson Harris. An information processing view of framing effects: The role of causal schemas in decision making. *Memory & Cognition*, 24(1):1–15, 1996.

[23] David Kaufman, Juli Murphy, Lori Erby, Kathy Hudson, and Joan Scott. Veterans' attitudes regarding a database for genomic research. *Genetics in Medicine*, 11(5):329, 2009.

[24] Amy A Lemke, Wendy A Wolf, Jennifer Hebert-Beirne, and Maureen E Smith. Public and biobank participant attitudes toward genetic research participation and data sharing. *Public Health Genomics*, 13(6):368–377, 2010.

[25] Sandeep Mishra, Margaux Gregson, and Martin L Lalumiere. Framing effects and risk-sensitive decision making. *British Journal of Psychology*, 103(1):83–97, 2012.

[26] ACMG Board of Directors. Laboratory and clinical genomic data sharing is crucial to improving genetic health care: a position statement of the american college of medical genetics and genomics. *Genetics in Medicine*, 19(7):721–722, 2017.

[27] Emily A Rauscher, Colin Hesse, Stacey Miller, Whitney Ford, and Erin L Youngs. Privacy and family communication about genetic cancer risk: Investigating factors promoting women's disclosure decisions. *Journal of Family Communication*, 15(4):368–386, 2015.

[28] Saskia C Sanderson, Kyle B Brothers, Nathaniel D Mercaldo, Ellen Wright Clayton, Armand H Matheny Antommaria, Sharon A Aufox, Murray H Brilliant, Diego Campos, David S Carrell, John Connolly, et al. Public attitudes toward consent and data sharing in biobank research: a large multi-site experimental survey in the us. *The American Journal of Human Genetics*, 100(3):414–427, 2017.

[29] Erwin L van Dijk, Yan Jaszczyszyn, Delphine Naquin, and Claude Thermes. The third revolution in sequencing technology. *Trends in Genetics*, 34(9):666–681, 2018.

[30] Lizzie Wade. To overcome decades of mistrust, a workshop aims to train indigenous researchers to be their own genome experts. *Science*, 9(27), 2018.