Submitted to Econometrica

THE EQUILIBRIUM EFFECTS OF INFORMATION DELETION: EVIDENCE FROM CONSUMER CREDIT MARKETS

Andres Liberman, Christopher A. Neilson, Luis Opazo and Seth Zimmerman[1]

Abstract

This paper uses a large-scale natural experiment to study the equilibrium effects of restricting information provision in credit markets. In 2012, Chilean credit bureaus were forced to stop reporting defaults for 21% of the adult population. Using panel data on the universe of bank transactions in Chile combined with the deleted registry information, we implement machine learning techniques to measure changes in the predictions lenders can make about default rates following deletion. Using a difference-in-differences design, we show that individuals exposed to increases in predicted default reduce borrowing by 6.4% following deletion, while those exposed to decreases raise borrowing by 11.8%. In aggregate, deletion reduces borrowing by 3.5%. Taking the difference-in-difference estimates as inputs into a model of borrowing under adverse selection, we find that deletion reduces surplus under a variety of assumptions about lenders' pricing strategies.

Keywords: Asymmetric Information, Consumer Credit, Pooling Equilibrium, Information Deletion, Equilibrium Effects, Credit Registry Information.

[1] We thank Andrew Hertzberg, Amir Kermani, Neale Mahoney, Holger Mueller, Christopher Palmer, Philipp Schnabl, Johannes Stroebel, and numerous seminar participants for comments and suggestions. Sean Hyland, Franco Calle and Jordan Rosenthal-Kay provided excellent research assistance. This research was funded in part by the Fama-Miller Center for Research in Finance and the Richard N. Rosett Faculty Fellowship at the University of Chicago Booth School of Business. We thank Sinacofi for providing the data.

1. INTRODUCTION

Many countries have institutions that limit the information available to consumer lenders. For example, in 2007, over 90% of countries with credit bureaus also had provisions that erased defaults after set periods of time (Elul and Gottardi, 2015). Other forms of information limits include restrictions on the types of past borrowing outcomes and demographic variables that can be used to inform future lending decisions, and one-time purges of default records. The stated motivation for these policies is often that allowing lenders access to certain kinds of information unfairly reduces borrowing opportunities for individuals with past defaults, who may be drawn disproportionately from disadvantaged groups or have suffered from a negative past shock such as a natural disaster, an economic downturn, or a health event (Miller, 2003; Steinberg, 2014).

Several recent empirical studies confirm that limits to public credit information increases borrowing for the direct beneficiaries of those limits.[1] But this gain may come at a cost to the non-direct beneficiaries with whom direct beneficiaries are pooled. The aggregate and distributional effects of limits to credit information depend on the tradeoff between these two groups, which is hard to evaluate empirically. The empirical challenge is to construct a plausible counterfactual for the evolution of credit outcomes for non-direct beneficiaries, whose information is unchanged by the limits to credit information, and who would be affected only as a consequence of lenders' response in equilibrium.

This paper exploits a large-scale, country-wide policy change to evaluate the effects of deleting credit information on consumer credit markets. In February 2012, the Chilean Congress passed Law 20,575 (henceforth, the

[1] See Musto (2004),Brown and Zehnder (2007), González-Uribe and Osorio (2014), Bos and Nakamura (2014), Herkenhoff, Phillips, and Cohen-Cole (2016),Liberman (2016), and Dobbie, Goldsmith-Pinkham, Mahoney, and Song (2016).

“policy change”), which forced all credit bureaus operating in the country to stop reporting individual-level information on defaults. The policy change affected information for all individuals whose defaults as of December 2011 added up to less than 2.5 million Chilean pesos (CLP; roughly USD $5,000), a group that made up 21% of all Chilean adults and 84% of all bank borrowers in default at the time of implementa tion. After the deletion, credit bureau information no longer distinguished individuals with deleted records from those with no defaults. The policy change was a one-time deletion and did not affect how subsequent defaults were recorded. Three years after the deletion, the count of individuals reported as in default in the credit bureau had nearly returned to its pre-deletion level and was still rising. We combine the policy change with administrative data that track bank outcomes and credit bureau data for the universe of bank borrowers in Chile.

We begin by using machine learning techniques to evaluate how the deletion policy affected banks’ ability to estimate borrowers’ expected probability of default. For each individual we train random forests to generate two sets of predictions. The first uses both bank borrowing data and credit bureau records, while the second uses only the bank borrowing data without the deleted credit bureau records. Eliminating credit bureau data reduces both in- and out-of-sample log likelihoods of observed values given predictions, and produces systematic overestimates of bank default probabilities for borrowers without defaults and underestimates for borrowers with defaults.

We define exposure to the deletion policy as percent increase in predicted bank default following deletion. As credit bureau non-defaulters outnumber credit bureau defaulters, exposure is positive (i.e., predicted bank defaults rise) for 61% of the population. Individuals with the highest exposure borrow lower amounts and are poorer, and resemble those with the lowest exposure except for the credit bureau default. In contrast, predicted bank default does not change after deletion for individuals who borrow large amounts with higher rates of bank default.

Using a difference-in-differences strategy, we show that the deletion policy increases borrowing for defaulters by 46% relative to non-defaulters. This finding is consistent with previous work on the effects of information deletion. However, it is uninformative about the *aggregate* effects of deletion because it reflects a combination of gains for defaulters and losses for non-defaulters. The empirical challenge in measuring aggregate effects is to construct counterfactuals for how consumer credit would have evolved for defaulters and non-defaulters in the absence of the policy change.

To estimate the aggregate effects of the deletion policy we exploit our measure of exposure to the policy chang. Intuitively, changes in banks' credit supply decisions are likely to be correlated with changes in predicted bank default rates, which we measure with exposure. We use snapshots of borrower and credit bureau data at six month intervals leading up to and including the December 2011 snapshot to identify groups of borrowers who would have been exposed to positive, negative, and zero changes in default predictions had deletion taken place at that time. We interact the predicted exposure variables with a dummy that equals one for the cohort exposed to the actual deletion policy (the December 2011 cohort) to estimate the effects of deletion in the positive- and negative-exposure group relative to the zero-exposure group. This exercise recovers the effects of deletion on borrowing in aggregate under the assumptions that, a) borrowing trends in the positive, negative, and zero exposure groups would have evolved in parallel in the absence of the policy, and b) that the policy does not affect borrowing levels in the zero-exposure group.

We find that quantities borrowed by the negative- and positive-exposure groups move in parallel to the zero exposure group during the pre-deletion period. Following deletion, borrowing jumps up by 11.7% for the group exposed to decreases in predicted default and falls by 6.4% for the group exposed to increases in predicted default. As lenders' predictions of default fall by 29% in the former group and rise by 22% in the latter, the elasticities of lending to predicted default equal -0.40 and -0.29 in the positive and negative exposure groups, respectively. Because more borrowers are exposed to increases in predicted bank default than to decreases, the aggregate effect of deletion across the two groups was to reduce borrowing by 3.5%, about $40 million USD over a six month period. The decline in borrowing is larger as a share of borrowing for lower-income borrower. The decrease in aggregate borrowing is consistent with a credit market with asymmetric information (Akerlof, 1970; Jaffee and Russell, 1976; Stiglitz and Weiss, 1981), and rules out that the policy merely induces redistribution across borrowers in the presence of constrained banks.

We evaluate the assumption that borrowing is unchanged for the zero-exposure group using a supplemental difference-in-differences analysis. We compare borrowing for defaulters in the zero-exposure group above the deletion cutoff–whose information was not deleted–to borrowing for below-threshold borrowers in the zero-exposure group–whose information was deleted. We find that deletion did not affect borrowing for the individuals in the zero-exposure group around the cutoff. In contrast, as expected, negative exposure borrowers below the threshold increase their borrowing significantly after the policy change relative to those above the threshold.[2]

To study the effects of the deletion policy on total surplus, we use a simple framework that takes as a baseline an unraveling model in the style of Akerlof (1970) and Einav, Finkelstein, and Cullen (2010). In the model, the effect of deletion on total surplus is ambiguous and depends on the demand and cost curves for high- and low-cost borrowers. We use the estimates from

[2] There are no positive exposure borrowers with defaults close to the policy threshold, because individuals near the policy threshold are in default.

our difference-in-differences analysis to construct these curves, mapping borrowers with negative exposure to the high-cost market and borrowers with positive exposure to the low-cost market. In a baseline scenario with average cost pricing we find that pooling increases total surplus losses from adverse selection by 66% relative to the no-pooling equilibrium –a result that holds qualitatively over a wide range of possible markups over rates. Because deletion may have dynamic welfare effects or welfare effects outside of the credit markets, we view our findings as measures of the costs of providing insurance and benefits outside the credit market.[3]

In the final section of the paper, we use our procedure to study the effects of two counterfactual policies that limit information available to lenders: deleting bank default records in addition to credit bureau default records, and deleting information on gender (Munnell, Tootell, Browne, and McEneaney, 1996; Blanchflower, Levine, and Zimmerman, 2003; Pope and Sydnor, 2011). Deleting additional default information increases the spread of changes in predicted bank default, with bigger gains for winners and, larger losses for losers than in the policy as implemented. Deleting information on gender increases predicted bank default disproportionately for women. The common theme is that the costs of deletion fall mostly on individuals observably similar to the intended beneficiaries.

This paper contributes to a broader literature on the empirics of asymmetric information. Our finding that deleting information reduces overall borrowing and that costs fall most heavily on non-defaulters who resemble defaulters is similar to Agan and Starr (2017), which shows that restricting information on criminal records in job applications reduces callback rates

[3] For example, periodic information deletion may help insure against the ex ante ‘reclassification’ risk of defaulting and losing access to credit markets (Handel, Hendel, and Whinston, 2015), or may induce externalities in labor markets (Bos, Breza, and Liberman, 2018; Herkenhoff, Phillips, and Cohen-Cole, 2016; Dobbie, Goldsmith-Pinkham, Mahoney, and Song, 2016). See also Clifford and Shoag (2016), Bartik and Nelson (2016), Cortes, Glover, and Tasci (2016), and Kovbasyuk and Spagnolo (2018).

for black applicants. We show how a machine learning approach can identify individuals affected by deletion policies, and, develop a framework that can be used to evaluate welfare effects.

We also contribute to a literature that uses machine learning to explore treatment effect heterogeneity given access to many mediating variables (Athey and Imbens, 2016; Athey and Wagner, 2017) and to generate counterfactuals that allow for causal inference where no credible experiment exists (Burlig, Knittel, Rapson, Reguant, and Wolfram, 2017).[4] In contrast to this work, we focus on measures of predicted average costs that are theoretically-motivated as the key determinant of heterogeneous treatment effects. This reduces the set of causal parameters required to apply our approach in other settings from a potentially large number of heterogeneous effects defined across interactions of mediator variables to a single set of elasticities. Our approach complements studies of how 'big data' is increasingly prevalent in credit markets and other settings (Petersen and Rajan, 2002; Einav and Levin, 2014).

2. EMPIRICAL SETTING

2.1. *Formal consumer credit and credit information in Chile*

In Chile, formal consumer credit is supplied by banks and by other non-bank financial intermediaries, most notably department stores. There were 23 banks operating in Chile as of December 2011, including one state-owned and 11 foreign-owned institutions, which had issued approximately $23 billion in non-housing consumer credit (i.e., credit cards, overdraft credit lines,

[4] See Varian (2016) or Mullainathan and Spiess (2017) for a review. Several other papers employ machine learning techniques to study credit markets. These include Huang, Chen, and Wang (2007), Khandani, Kim, and Lo (2010) and Fuster, Goldsmith-Pinkham, Ramadorai, and Walther (2017). These papers focus on using machine learning techniques to improve cost prediction. In contrast, we use ML techniques to study the effects of actual and counterfactual policy changes on borrowing.

and unsecured term loans).[5] At the same time, the 9 largest non-banking lenders (all department stores) had a total consumer credit portfolio of approximately $5 billion. Although banks issue more credit, department stores lend to more borrowers (14.7 million active non-bank credit cards, of which 5.4 million recorded a transaction during that month, versus 3.8 million consumer credit bank borrowers).[6]

Banks (and non-bank lenders) rely on defaults reported in the credit bureau to run credit checks of potential borrowers (Cowan and De Gregorio, 2003; Liberman, 2016). Defaults reported to the credit bureau include bank and non-bank debt, as well as other obligations such as bounced checks and utility bills. Importantly, banks are required by law to disclose their borrowers' outstanding balance and defaults to the banking regulator (SBIF), which then makes this information available exclusively to banks. As a result, banks may learn a borrower's total bank debt and bank defaults, but may only observe reported defaults from non-banks (i.e., cannot access non-bank debt balances). In turn, non-banks can only learn an individuals' bank and non-bank defaults from the credit bureau, not the level of bank or non-bank consumer credit.

2.2. *The policy change*

In early 2012, the Chilean Congress passed Law 20,575 to regulate credit information.[7] The bill included a one-time "clean slate" provision by which credit bureaus would stop sharing information on individuals' delinquencies that were reported in December 2011 or earlier. This provision affected only borrowers whose total defaults, including bank and non-bank debts, added up to at most 2.5 million pesos. According to press reports, the provision was

[5] All information in this paragraph is publicly available through the local banking regulator's website, www.sbif.cl.

[6] Chile's population is approximately 17 million.

[7] See http://www.leychile.cl/Navegar?idNorma=1037366.

a way to alleviate alleged negative consequences of the February 2010 earthquake, which had caused large damage to property and ostensibly forced a number of individuals into financial distress. The Chilean Congress had already enacted a similar law that forced credit bureaus to stop reporting information on past defaults in 2002. Nevertheless, this new "clean-slate" was marketed as a one-time change, and indeed, all new defaults incurred after December 2011 were subsequently subject to the regular treatment and reported by credit bureaus.

Following the passage and implementation on February 2012 of Law 20,575, credit bureaus stopped sharing information on defaults for roughly 2.8 million individuals, approximately 21% of the 13 million Chileans older than 15 years old.[8] In effect, this means that individuals who were in default on any bank or non-bank credit as of December 2011 for a consolidated amount below 2.5 million pesos began to appear as having no defaults. This is shown in the left panel of Figure 1, where we plot the time series of the number of individuals in our data with any positive default reported through credit bureaus as of the last day of each semester (ending in June or December).[9] The left panel of the figure shows a large reduction in the number of individuals with any defaults as of June 2012, after the policy change, relative to December 2011.[10] Interestingly, the figure shows a sharp increase in the number of affected individuals in the following semesters. This is consistent with the fact that the policy was a one-time change, as future defaults were recorded and reported by credit bureaus, as well as with the fact that many individuals whose defaults were no longer reported did default on new obligations.

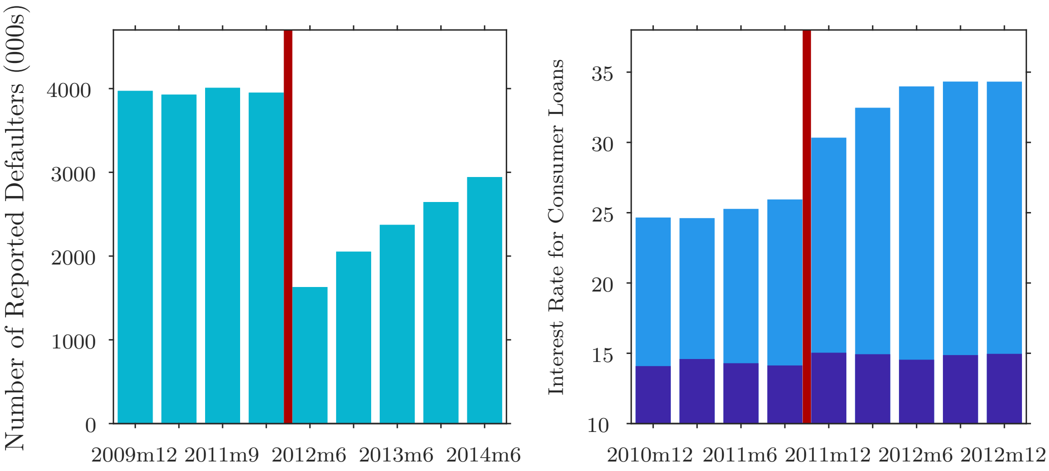
[8] Figure taken from press reports of the "Primer Informe Trimestral de Deuda Personal", U. San Sebastian.

[9] Due to data constraints, our data is limited to individuals who were present in the regulatory banking dataset prior to the passage of Law 20,575.

[10] There is no evidence of an aggregate increase in defaults following the February 2010 earthquake.

The policy change modified the information that lenders, bank and non-bank, could obtain on defaults at other lenders. After the policy change, non-bank lenders could no longer verify any type of defaults, while banks could not observe whether individuals had defaulted on non-bank debt. However, banks could still verify whether an individual had bank defaults because the banking regulator's data was not subject to the policy change. Thus, the policy change induced a sharp information asymmetry between the banking industry as a whole and its borrowers, rather than creating asymmetries in the information available to each bank with respect to its borrowers.

Figure 1: Reported Accounts in Default and Consumer Loan Interest Rates



In the left panel, each bar in lighter 🟦 represents the count of individuals in the credit registry with positive default values at six month intervals. The vertical line represents the implementation of the registry deletion policy. In the right panel we show average interest rates by quarter for small consumer loans in lighter blue 🟦 and large consumer loans in darker blue 🟪. Information on rates obtained from website of Superintendencia de Bancos e Instituciones Financieras, `www.sbif.cl`.

The median interest rate charged to small borrowers rose following deletion. The right panel of Figure 1 plots median interest rates for small and large consumer loans before and after the deletion. We observe a 5.3 percentage point increase in rates in the small loan market, a 20% rise from a base of 26%. Rates continue to rise following the policy change, reaching almost

35% (30% above the base pre-policy rate) by the fourth quarter following implementation. We do not observe changes in rates for larger borrowing amounts, which suggests that the effects we see are not driven by coincident changes in other determinants of borrowing rates. We show below that on average most new borrowing is done by borrowers with no defaults. This means that the median new loan can be thought of as belonging to this market.

2.3. *Data and summary statistics*

We obtain from Sinacofi, a privately owned Chilean credit bureau, individual-level panel data at the monthly level on the debt holdings and repayment status for the universe of bank borrowers in Chile from April 2009 until 2014. Sinacofi has access to the banking data that are not available to other credit bureaus because Sinacofi's only clients are banks. Sinacofi merged the data to measures of consolidated defaults from the credit registry. We observe registry data at six month intervals, in June and December of each year. As is typical in most empirical research on consumer credit, microdata do not include interest rates or other contract terms.

We use these data to build a panel dataset that links snapshots of defaults as reported to the credit bureau to borrowing outcomes. We use the six credit bureau snapshots from December 2009 through December 2011. We link each snapshot to bank borrowing and default outcomes over the six month period beginning two months after the snapshot (i.e., the six month interval beginning in February for the December snapshots, and the six-month interval beginning in August for the June snapshot). This alignment corresponds to the timing of the deletion policy, which took place in February 2012 based on the December 2011 credit bureau default records.

Table I reports summary statistics for these data. The first column is the full sample, which includes all individuals who show up in the borrowing data. There are 23 million person-time period observations from 5.6 million individuals in the dataset. Around 37% of borrowers in our dataset have a positive value of credit bureau defaults, with an average value in default of \$554,500 CLP, and 31% of the population, or 84% of all defaulters, have a default amount strictly between 0 and \$2.5 million CLP, and are eligible for deletion. We observe deletion for 29% of all individuals in the December 2011 cohort. The two percent gap between our calculated deletion eligibility rate and observed deletion rate is due to rare default types that are not included in the consolidated measure we observe. Conditional on eligiblity for deletion, the average consolidated amount in default is \$172,250 CLP.[11]

The average bank debt balance for consumers is \$7.8 million CLP. Unsecured consumer lending accounts for 28% of all debt, for an average of \$2.2 million CLP. Mortgage debt accounts for the majority of the remainder. The average bank default balance (defined as debt on which payments are at least 90 days overdue) across all borrowers is \$338,090 CLP, or 12% of the overall debt balance. For borrowers eligible for deletion of defaults, this average is \$147,460. Comparing bank default balances to credit bureau default balances shows that deletion eliminates banks' access to 15% $(= 100 \times (1 - 147/172))$ of the default amount among individuals whose balances in default falls below the deletion threshold.

We do not directly observe new borrowing or repayment. Thus, we define new consumer borrowing as any increase in an individual's consumer debt balance of at least 10% month over month. We define as the amount of the increase times an indicator for new borrowing times the amount of the increase. In the full sample, 30% of consumers take out at least one new consumer loan in the six month period following each credit snapshot. The average amount of new borrowing is \$184,000 CLP. We define new

[11] Figure 1 of the Online Appendix presents a histogram of the default amount as of December 2011 for all individuals and for individuals with positive defaults.

bank defaults analogously using borrowers' bank default balances. 17% of customers have a new bank default, with an average default amount of $37,000 CLP. In our analysis of the effects of information deletion we focus on new consumer borrowing as the outcome of interest as defaults are most costly to lenders for uncollateralized borrowing.

The average age in our sample is 44, and 44% of borrowers are female. Our data identify borrowers' socioeconomic status for 10% of individuals overall. These data, which were collected by banks, divide individuals into five groups by socioeconomic background. We use these data to generate predictions of socioeconomic status for all individuals in the sample using a machine learning approach.[12] In our empirical analysis, we split our sample by this predicted SES categorization. One strong predictor of SES classification is whether or not an individual has a home mortgage. We split our sample by this categorization as well.

The second column of Table I describes our main analysis sample. We focus on borrowers who have a positive debt balance six months prior to the credit snapshot and consolidated default of $2.5 million CLP or less, including zero values. This group accounts for 97% of individuals and 95% of observations. The restriction on debt balances allows us to define a consistent sample across time. Without it, the structure of our data generates spurious increases in mean borrowing over time. This occurs because individuals are included in our sample only if they borrow at some point between 2009 and 2014. An individual with a zero debt balance in 2009 must borrow in the future; otherwise, she would not be included in the data. Subsetting on individuals with positive debt balances at baseline addresses this issue.[13] The restriction to consolidated defaults of $2.5 million CLP or less lets us focus on the part of the credit market where available information changed.

[12] We describe this process in detail in the Online Appendix.

[13] An alternate approach would be to take the population of all Chileans, irrespective of borrowing, as the sample. We do not have access to data on non-borrowers.

Lenders were able to observe consolidated defaults above $2.5 million CLP both before and after the cutoff. Demographics and borrowing in the panel sample are similar to the full dataset.

TABLE I

Sample Description

	All	In Panel	In Panel, Positive Borrowing
Any registry default	0.37	0.33	0.14
Deletion eligible	0.31	0.33	0.14
Observed deletion	0.29	0.30	0.17
Registry default amt.	554.50	182.00	54.45
Reg. default amt | reg. <2.5m	172.25	182.00	54.45
Debt balance	7,768	7,675	13,075
Consumer borrowing balance	2,172	2,097	2,634
Have mortgage	0.19	0.19	0.24
Mortgage balance	4,343	4,387	8,192
Any bank default	0.17	0.14	0.03
Bank default amt.	338.09	155.81	31.06
Bank default amt | reg. <2.5m	147.46	155.81	31.06
Default amt./balance	0.12	0.09	0.01
New consumer borrowing	0.31	0.32	1.00
New consumer borrowing amt.	184	190	650
New bank default	0.08	0.08	0.05
New bank default amt.	36.57	27.28	14.55
Age	44.12	44.08	43.40
Female	0.44	0.45	0.45
Have SES	0.10	0.10	0.13
SES A	0.25	0.25	0.36
SES B	0.29	0.29	0.27
SES C	0.25	0.25	0.20
SES D & E	0.22	0.22	0.17
N of observations	23,001,337	21,769,213	4,593,511
N of clusters	330	330	330
N of individuals	5,577,605	5,433,403	2,314,786

Observations are at the person by half-year level. Data run from August 2009 through July 2012. Six-month credit bureau snapshots run from February-July and August-January. Borrowing outcomes from each six month interval are linked to credit bureau data from two months prior to the start of the interval (December and June, respectively). We refer to time periods by the bureau month. Columns define samples. 'All' column is all Chilean consumer bank borrowers. 'In panel' is the set of borrowers with a positive balance six months prior to a given month. 'In panel, positive borrowing' is the subset of borrowers who additionally have new borrowing in the snapshot – a 10% random sample of this subset defines our machine learning training set, which we exclude from the main panel. See text for details.

The third column of Table I describes the sample of individuals with positive borrowing. As we discuss in the next section, this is the sample we use for constructing cost predictions. They tend to be richer and have much lower current default balances relative to overall borrowing (0.01 vs 0.09 in the full panel). Their rates of future bank default are also somewhat lower (0.05 vs. 0.08 in the full panel).

3. MEASURING THE LOSS OF INFORMATION

Banks' prediction of individuals' future repayment is an important determinant of their credit supply decisions (Agarwal, Chomsisengphet, Mahoney, and Stroebel, 2018; Dobbie, Liberman, Paravisini, and Pathania, 2018). To form these predictions, banks rely on credit models that divide potential borrowers into groups based on observable characteristics and make predictions about future repayment within each group. We have access to borrowers' observable characteristics but do not observe banks' grouping choices. We implement a random forest algorithm and focus on a prediction of an indicator variable equal to one if a borrower adds to his default balance in the six month period following each registry snapshot.

The random forest repeatedly chooses sets of possible predictor variables at random and constructs a regression tree using those predictors. Each tree iteratively splits by the explanatory variables, choosing splits to maximize in-sample predictive power. The random forest obtains predictions by averaging over predictions from each tree. One way to think about this process in our context is as averaging over different guesses about which variables banks might use to classify borrowers. When predicting default outcomes, we focus on the sample of individuals who have new borrowing over that same period. We make this restriction because the goal of the exercise is to recover cost predictions for market participants.

We build each tree in our random forest by choosing variables at random from a set of 15 possible predictors. These consist of two lags (relative to the time of policy implementation) of new quarterly consumer borrowing, new quarterly total borrowing, consumer borrowing balance, secured debt balance, average cost, and available credit line, as well as a gender indicator.

For pre-policy predictions, the set of variables also includes the credit bureau default data. We set the number of trees in a forest to 150. Predictive power is not sensitive to other choices in this range. We choose other model parameters (how many variables to select for inclusion in each tree and the minimum number of observations in a terminal node in the tree) using a cross-validation procedure. For comparison, we also construct predictions using two alternate methods: a logistic LASSO and a näive Bayes classifier. See the Online Appendix for details on these approaches.

For each method, we construct two sets of predictions. The first set uses training data from the same registry cross-section as the outcome data. These predictions correspond to the best guess a lender can make about default outcomes using data available to them at the time of the loan. For this set of predictions, differences between predicted default with and without the default information depend on 1) differences in the average default rate in each submarket in the market equilibrium prior to the reform, 2) potentially time-varying shocks to credit demand, which move individuals with different covariate values along their cost curves, and 3) endogenous responses to the pooling policy (in the post-pooling time period).

Our objective is to isolate variation in predicted default due only to supply-side price shocks. Our second set of predictions helps us do this. This set of predictions uses training data from the December 2009 credit bureau default cross section to generate predictions for all other cross sections. Conditional on covariates, these predictions do not vary across cohorts in the remaining data and therefore do not reflect the effects of time-varying demand shocks. They use only data from before pooling took place, so they do not reflect endogenous reponses to information deletion.

We present separate estimates for predictors trained in the pre-period and those trained contemporaneously. The contemporaneous random forest predictions have in-sample (out-of-sample) log likelihood values of -0.173

(−0.295) when including registry information. Without registry information, these values fall to −0.177 (−0.305). The pre-period random forest predictions have slightly higher log likelihoods in both the training and testing sample, with a similar percentage decline from dropping registry information. Random forest predictions outperform the näive Bayes and logistic LASSO predictions.[14]

3.1. *The distribution of exposure to changes in predicted default*

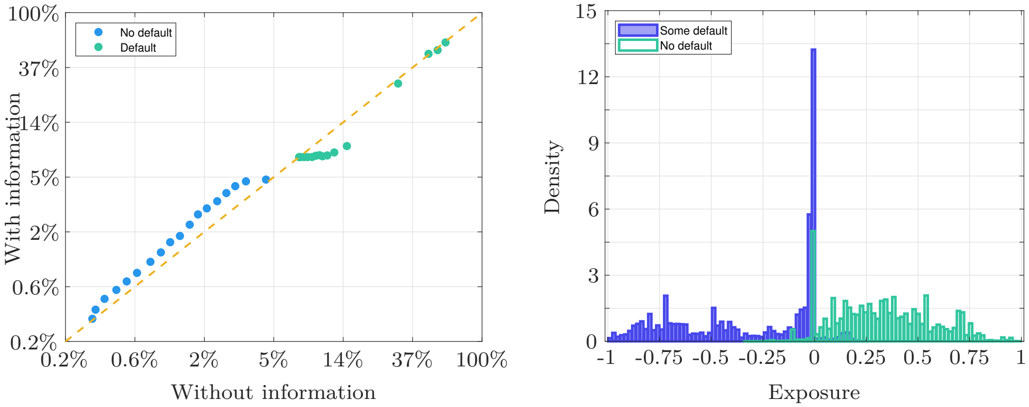
In addition to reducing explanatory power, deletion affects the distribution of bank default predictions across credit bureau defaulters and non-defaulters. Figure 2 describes these changes in two ways. The left panel shows the means of predictions made without default information within bins defined by values of the predictions that include default information. We split the sample by credit bureau default status. For individuals without defaults, deletion increases predicted default on average (points are above the 45-degree line). For individuals with defaults, deletion reduces default predictions (points are below the 45-degree line). Individuals with very high default probabilities do not seem affected, presumably due to the large amount of negative information available about their credit histories in addition to recent default.

The right panel of Figure 2 explores the distribution of changes in predicted values from deletion in more detail. For each individual, we define a measure of exposure to the deletion of information E_i as the percentage change in default prediction caused by deletion. For non-defaulters, predicted default rises for 89% of borrowers, with an average increase of 29%. For defaulters, predicted default falls for 95% of borrowers, with an average drop of 32%. The exposure distribution for defaulters is bimodal, with one

[14] Table 4 of the Online Appendix compares in- and out-of-sample log likelihood measures for the random forest to those from other prediction methods.

mode at zero and the other centered near a decline of 75%. More borrowers are non-defaulters than defaulters, so predicted bank defaults increase for a majority (63%) of borrowers in the market.

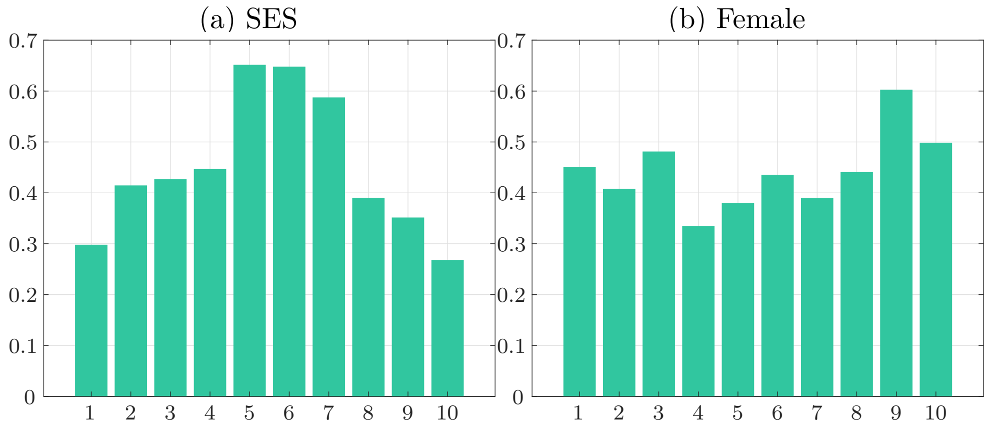
Figure 2: Predictions with and without registry data



Left panel shows the binned means of default predictions made with default registry data (horizontal axis, log scale) plotted against the binned means of default predictions made without default registry data (vertical axis, log scale). Predictions for the those with no prior default shown in blue 🟦 while those with positive prior default are shown in lighter green 🟩. The right panel shows a histogram of changes in predicted log bank default. Bars shown in darker blue 🟦 show exposure for defaulters while bars shown in lighter green 🟩 show exposure for non-defaulters.

Figure 3 plots binned means of indicators for coming from a high-SES background (left panel). The graph has an upside-down V shape, where about 25% of borrowers in the top and bottom deciles of the exposure distribution come from high-SES backgrounds, compared to a maximum of over 60% for individuals with exposure to slight increases in default predictions. Intuitively, borrowers who benefit most from the policy change, who see the largest drop in predicted default, are those who are difficult to distinguish from non-defaulters without access to the deleted information. In contrast, borrowers who are relatively unaffected by the policy are those for whom more accurate information about defaults is available outside of the deleted registry. The right panel shows a more stable pattern for a dummy for female borrowers across the distribution of exposure.

Figure 3: Borrower SES and gender by exposure to information deletion



Binned means of indicators by decile of exposure distribution for coming from a high-SES background in the left panel and for females in the right panel. Horizontal axis is log change in predicted default rate from deletion. ML predictions come from contemporaneous training dataset.

4. THE EFFECTS OF INFORMATION DELETION ON CREDIT OUTCOMES

This section reports our main empirical analysis to estimate the effects of the deletion policy on credit markets. We start by reporting the change in borrowing outcomes–default predictions and consumer borrowing– for individuals with deleted credit bureau records relative to individuals whose records were not deleted. Then, we present our main empirical strategy to isolate the causal effects of deletion on consumer credit outcomes that exploits the change in predictions of bank default induced by the policy, as described in the previous section.

4.1. *The effects of deletion for defaulters relative to non-defaulters*

We first report how borrowing and predicted bank default change for individuals with deleted credit bureau default records relative to individuals without deleted records using the full sample of borrower data in each credit bureau snapshot. To compare the evolution of borrowing and predicted default for both groups over time, we estimate difference-in-differences specifications that interact the individual's cohort relative to deletion with $D(\text{Positive Default})_{it}$, which is an indicator variable for a positive default on the credit bureau snapshot:

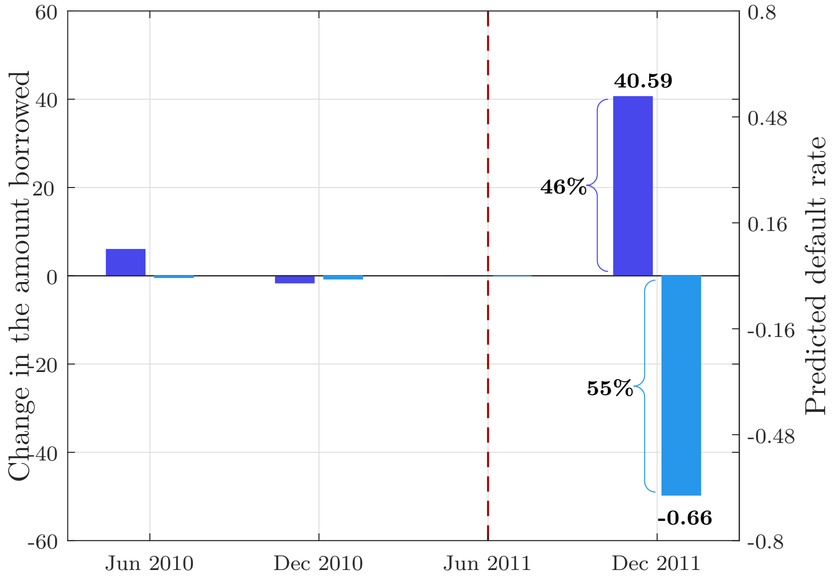
$$(1) \qquad Y_{ic} = \gamma_c + \gamma_c D(\text{Positive Default})_{ic} + X_{ic}\Psi_c + e_{gt},$$

where Y_{ic} is borrowing for individual i in cohort c, γ_c are cohort fixed effects, and X_{ic} are a set of individual covariates that include age, gender, and lagged borrowing and default outcomes. We define Y_{ic} as new consumer borrowing in the six month period that starts two months after each credit bureau snapshot. This definition of the outcome variable ensures that we measure new consumer borrowing for the cohort affected by the policy change, the December 2011 cohort, after the law is implemented in February 2012.

The coefficients of interest of regression (1) are the γ_c, which capture the difference in new consumer credit for borrowers in default relative to borrowers not in default in each cohort, and relative to the omitted cohort. In order to interpret the coefficient for the December 2011 cohort as the causal effect of the policy deletion on the difference in borrowing for defaulters and non-defaulters, we assume that new consumer borrowing for both groups would remain in parallel trends absent the policy change, which we verify with pre-trends.

Figure 4 plots the estimated parameter capturing the difference across groups of those with and without positive default. When the dependent variable is new consumer borrowing (on the left axis), we can see that the differences in borrowing are steady in the year leading up to deletion, validating the identification assumption. In the six months following deletion, borrowing for defaulters rises 46% relative to borrowing for non-defaulters (an increase of $41,000 CLP with a base-period borrowing of $88,000 CLP for defaulters). Figure 4 also shows the estimated parameter capturing the difference across groups when the dependent variable is the log of predicted bank default in the next 6 months (on the right axis). The log difference in bank default prediction is steady in the year leading up to deletion, then falls by 0.66 after deletion, corresponding to a 52% decline in banks' default expectations for defaulters relative to non-defaulters.

Figure 4: Effects of registry deletion on defaulters relative to non-defaulters



Difference-in-difference estimates of the effects of prior default on predicted default rate shown in darker blue ■ (left axis) and the change in the observed borrowing in lighter blue ■ (right axis) using equation 2. Both post policy coefficents are significant at 95% confidence levels. See text for details.

Our findings imply that the deletion of credit bureau defaults raises borrowing for the beneficiaries of deletion *relative* to non-beneficiaries. However, this estimate reflects a combination of gains for defaulters and losses for non-defaulters as banks have a difficult time differentiating among borrowers and cannot be interpreted as a causal estimate of the aggregate effect of the deletion of credit information on consumer borrowing. Next, we present our empirical strategy that makes use of changes to banks' default predictions in order to estimate the causal effects of the deletion of information.

4.2. *The causal effects of deletion on consumer borrowing*

We isolate the effects of changes in lenders' predictions about future bank default on borrowing outcomes using a difference-in-differences approach that exploits our predictions of exposure to the policy change described in Section 3. Intuitively, we compare changes in borrowing outcomes before and after deletion for individuals exposed to increases (and decreases) in beliefs about future bank default to those for individuals with near-zero exposure. A crucial assumption we make is that banks' credit supply decisions are correlated with expected default. Although this measure of costs– defaults– is not comprehensive, it is likely to be correlated with banks' supply decisions and ex ante profits. For example, Dobbie, Liberman, Paravisini, and Pathania (2018) show that banks focus more on default than other measures of costs due to agency concerns with loan officers.

Consider a sample of individuals who are either not exposed to changes in lender beliefs to deletion, or who are exposed to increases (decreases) in predicted bank default. Within this sample, we estimate specifications of the form:

$$Y_{ic} = \gamma_c + \tau_c D_{ic} + X_{ic}\Psi_c + e_{ic}. \tag{2}$$

Here, D_{ic} is an indicator equal to one if an individual is in the group exposed to increased (decreased) predicted bank default. The coefficients of interest are the τ_c, which capture cohort-specific estimates of the effects of exposure to increases in bank default predictions on borrowing. γ_c are cohort-fixed effects and X_{ic} are borrower-level controls. We normalize τ_c to be zero in the cohort immediateley prior to deletion. If deletion reduces borrowing for exposed individuals, we expect τ_c to be flat in the cohorts leading up to treatment, and then to become negative in the deletion cohort.

We measure exposure using random forest predictions trained in the December 2009 pre-period, as described in section 3. We split borrowers into three groups according to the change in predicted default: the "positive-exposure market", defined as individuals for whom default predictions rise by at least 15% following deletion, the "negative-exposure market," defined as individuals for whom default predictions fall by at least 15%, and the "zero group," defined as individuals for whom default predictions change by less than 15% in either direction. Our findings are robust to changing this threshold value.[15] When computing exposure, we winsorize values in the bottom 5% of the predicted distributions of default with and without registry data to avoid classifying very small differences in predicted default levels as large log differences. Our findings are not affected by modifying the winsorization threshold slightly.

Most borrowers are exposed to increases in predicted default from deletion: 53% of observations fall into the positive-exposure category, compared to 32% in the zero-change group and 16% in the negative-exposure group. Almost all borrowers in the negative-exposure group have bank defaults, while almost no borrowers in the positive-exposure group do.

This type of specification can recover the total effect of deletion on borrowing under two assumptions. The first is the standard difference-in-differences assumption that borrowing in the non-zero exposure groups follows parallel trends to the zero exposure group. We evaluate this assumption by looking at pre-trends in the τ_c. The second assumption is that deletion of credit bureau defaults does not affect borrowing outcomes for individuals in the zero-exposure group. If the deletion raised (lowered) borrowing in the zero-

[15] We have estimated alternate specifications that vary the threshold between 5% and 25%; results available upon request.

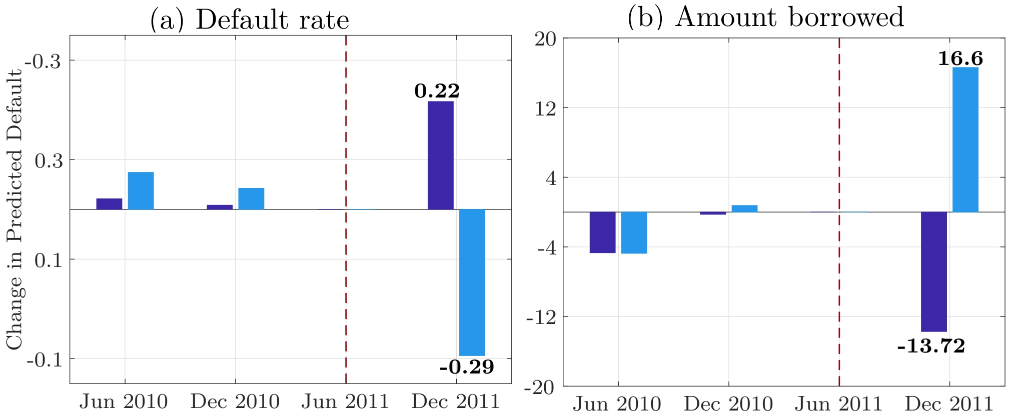
exposure group, our estimates will understate (overstate) the gains in borrowing attributable to deletion. We revisit this assumption below using a supplementary difference-in-differences approach.

Statistical inference is not straightforward in this setting. We would like to allow for correlation in error terms within the categories that banks use to estimate default, but we do not observe what these categories are. We use an auxiliary machine learning step to identify interactions of covariates within which individuals have similar expected default (i.e., each of these interactions identifies smaller "markets" where borrowers look similar to lenders). We then cluster standard errors in our regressions within groups defined by these interactions. There are 330 such groups in the full sample. Inference is robust to changes in the coarseness of these groupings.

Figure 5 and Table II report estimates of equation 2. These estimates recover effects for borrowers exposed to positive and negative shocks to bank default predictions relative to the group where bank default predictions do not change following deletion. Banks' expectations for both groups are flat in the year leading up to deletion.

At the time of deletion, log bank default predictions rise by 0.22 in the positive exposure group and fall by 0.29 in the negative exposure group. Pre-trends in borrowing are also flat for both groups in the year leading up to deletion. Following deletion, borrowing falls by \$14,000 CLP in the positive exposure group, equal to 6.4% of pre-period mean for that group. Borrowing rises by \$17,000 CLP for the negative exposure group, equal to 11.8% of the pre-deletion mean. The implied elasticity of borrowing with respect to changes in default predictions is -0.29 (-0.40) in the positive (negative) exposure group.

Figure 5: Effects of registry deletion by changes in predicted default


(a) Default rate

(b) Amount borrowed

Note: Figure presents results of difference-in-difference estimates of the effects of exposure to changes in predicted default rate on predicted default rate (left panel) and new borrowing (right panel) using equation 2. Each panel splits the sample into individuals with positive change in predicted default (high exposure) shown in darker blue ■ and negative effects in predicted default (low exposure) shown in lighter blue ■. Effects for each group are measured relative to the omitted category of no exposure to changes in predicted default, defined as the bottom fifteen percent of the distribution of the absolute value of predicted default changes. Standard errors clustered at market level. See Table II and text for details.

These estimates indicate that the net effect of deletion was to reduce borrowing. The group exposed to increases in predicted default consists of 2.1 million individuals. At an average loss of $14,000 CLP per person, the total loss is just under $30 billion CLP, or $60 million USD at an exchange rate of 500 CLP per dollar. The group exposed to decreases in predicted default consists of 608,000 individuals, with an average gain of $17,000 CLP per person and a total gain of $10 billion CLP or $20 million USD. The net effect of deletion across the two markets was thus to reduce borrowing by $20 billion CLP, or 3.5% of the total borrowing across the two groups.[16] To the extent the goal of deletion policy was to increase access to credit, it appears to have been counterproductive.

[16] This aggregate drop in credit is also noted by Kulkarni, Truffa, and Iberti (2018).

TABLE II

Difference in differences by default and exposure

	Positive exposure		*Negative exposure*	
	Predicted Defaults	New Borrowing	Predicted Defaults	New Borrowing
Jun. 2010	0.02	-4.67^{+}	0.07	-4.74^{*}
	(0.03)	(2.81)	(0.08)	(2.30)
Dec. 2010	0.01	-0.25	0.04	0.75
	(0.03)	(3.25)	(0.07)	(2.59)
Dec. 2011	0.22^{***}	-13.72^{***}	-0.29^{***}	16.60^{***}
	(0.04)	(3.83)	(0.06)	(3.72)
Elasticity		-0.29		-0.40
Dep. Var. Base Period Mean	0.04	215.28	0.10	140.98
N Clusters	303	303	282	285
N Obs.	2,910,733	13,093,725	1,273,371	7,493,968
N Individuals	1,836,294	4,363,940	986,205	3,212,628
N Exposed Individuals	505,295	2,132,055	84,746	608,229

Significance: $^{+}$ 0.10 * 0.05 ** 0.01 *** 0.001. Difference and difference estimates from equation 2. The first two columns report the difference-in-difference estimated effect of deletion on outcome variables listed in column headers, while the third and fourth estimate the dif-in-dif effect on the different exposure-defined markets. Sample in specifications where cost is an outcome conditions on positive borrowing (see text for details). We take the log of 'Predicted Default' for estimation but report the base period mean in levels. 'Elasticity' is borrowing effect scaled by base period outcome mean and predicted default effect. 'N exposed individuals' reports the number of individuals not in the 0 group included in the regression sample in the treatment period. Since some individuals appear in multiple snapshots we report both individuals and observations. Standard errors clustered at market level. See text for details.

Online Appendix Tables 2 and 3 repeats the analysis from Table II, subsetting by whether borrowers have a mortgage at baseline and by our predicted measure of socioeconomic status, respectively. The common theme is that the effects of deletion are largest for the low-SES borrowers who are most exposed to changes in predicted costs.

4.2.1. *Comparison to no-deletion group*

To support the assumption of no effect on the zero-exposure group, we test for differential changes in new consumer borrowing for individuals whose credit bureau defaults add up to less than 2.5 million CLP, who were exposed to the policy change, relative to individuals whose defaults add up to more (or equal) than 2.5 million CLP, who were not exposed to the policy change.

Intuitively, if the zero-group is indeed unaffected by the deletion policy, then a comparison of individuals with small changes to their default predictions below the policy cutoff, who were affected by the deletion, and above the policy cutoff, who were unaffected, will show no difference in borrowing outcomes after the policy change.

To control non-parametrically for differences in new borrowing along the distribution of amount in default, we restrict our analysis to a bandwidth of 250 thousand CLP around the policy cutoff.[17] We compute this change in new borrowing for the three cohorts prior to the policy change (June 2010, December 2010, and June 2011) and the cohort exposed to the policy change (December 2011).

For each cohort we divide the sample in two groups defined by our machine learning predictions: negative-exposure individuals, for whom predicted default drops by more than 15%, and the zero-exposure group. There are no individuals exposed to an increase in predicted default in this sample of individuals, as these are all individuals who already are in default at relatively high amounts.[18] We run the following specification differentially for the two groups:

$$(3) \qquad Y_{ic} = \gamma_c + \tau_c \times 1[\mathit{Default}_{ic} < 2,500,000] + e_{ic},$$

where, again, Y_{ic} is borrowing for individual i in cohort c and the γ_c are cohort fixed effects. $1[\mathit{Default}_{ic} < 2,500,000]$ is an indicator equal to one if total credit bureau defaults for individual i in cohort c add up to less than

[17] Our findings are robust to widening or narrowing this bandwidth, although standard errors grow due to small sample sizes at very narrow bandwidths. We obtain near-identical findings in RD-DD specifications that allow for separate linear trends in default amount above and below the cutoff value in each cohort relative to policy change. These results are available upon request.

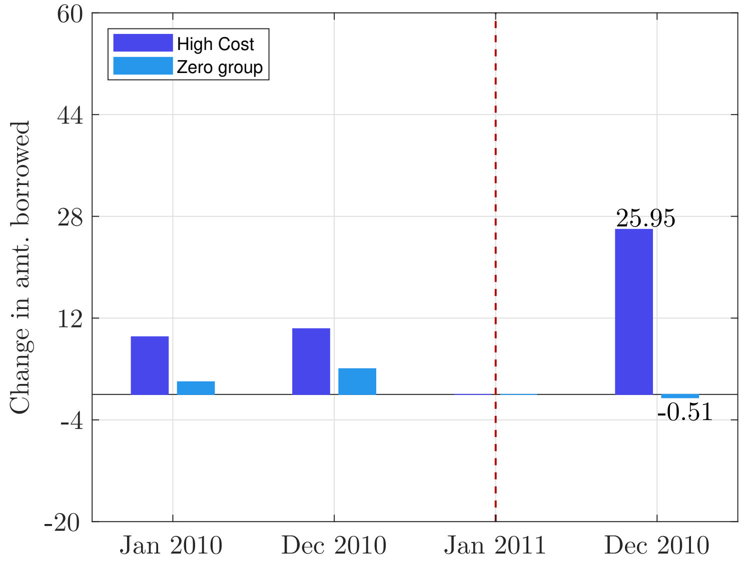
[18] To compute predicted default for the above-threshold group under the information deletion policy we apply the predicted values from the machine learning exercise described above based on observable covariates X_{ic}.

2.5 million CLP. The τ_c are the effects of interest, capturing how borrowing changes after registry deletion in 2011 for individuals whose amount in default is less than the policy cutoff of 2.5 million CLP.

This test recovers the causal effect of the policy change for the zero-exposure and negative-exposure groups under the assumption of no differential trends for individuals above and below the cutoff, which we examine visually with pre-trends. If our assumption that deletion does not affect borrowing for the zero-exposure group is correct, we should see no change in outcomes for this group following deletion. Moreover, an increase in borrowing for the negative-exposure group right below the cutoff would help make the zero-group test more compelling by showing that the deletion policy and our measures of exposure to that policy are good predictors of outcomes not just overall but also within the subgroup of relatively large defaulters.

We present the findings in Figure 6. The coefficients of interest of equation (3) for the zero-group are indistinguishable from zero before the policy change, indicating no pre-trends. They are also indistinguishable from zero after the policy change, which is consistent with the identification assumption for our main analysis. The graph also shows a large increase in borrowing for high-default individuals, exposed to decreases in predicted default, whose defaults are less than the 2.5 million CLP cutoff after the policy change. This rules out that the absence of an effect for the zero-group after the policy change is driven by a lack of power to identify any effects of the policy change among high-default individuals and is consistent with the main findings in this paper.

Figure 6: Effects of registry deletion at the policy cutoff



Difference-in-difference estimates for effects of the policy change at the policy cutoff of 2.5 million pesos using equation 3 for the exposure-defined 'zero group' in lighter blue 🟦 and 'negative exposure' in darker blue 🟦. These estimates compare new borrowing for individuals whose defaults are less than the cutoff relative to those whose defaults are higher than the cutoff, before and after the policy change, for the low exposure and zero groups. See text for details.

4.3. *Additional evidence: borrowing from non-banks*

The effects of deletion on aggregate borrowing could be reduced if individuals subject to higher prices for bank credit shift towards non-bank borrowing. The largest non-bank consumer lenders in Chile are department stores that issue credit cards. We explore how borrowing changed at these institutions using publicly-available aggregate data on retail credit card lending provided by SBIF. Online Appendix Figures 3a, 3b, and 3c show no distinct breaks in the total stock of retail credit cards, the number of retail credit cards used, or the amount transacted at the time of deletion.

These findings are consistent with the hypothesis that deletion reduced aggregate borrowing. Deletion effects in the retailer-issued credit card market may be smaller than in the consumer bank lending market because low-risk individuals are very unlikely to borrow in that market both before and after deletion. Median interest rates for retailer credit card lending are 75% higher than for non credit-card consumer bank lending just before deletion (45% vs. 26% in November 2011) and remain higher following deletion (e.g. 45% vs. 31% in February 2012).[19] That few individuals subsitute from consumer credit to credit card borrowing is consistent with the observation that prices remained lower in the consumer credit market following the deletion.

In fact, the deletion may have induced a larger effect on non-defaulters among non-banks than banks. While banks continued to observe bank defaults (at all other banks) following deletion, the deleted credit bureau information was the only default information available to non-bank lenders. Because there is no micro-level data for non-bank lenders, we cannot directly calculate how exposure to the policy affects non-bank lending, but our results for bank lending suggest there may be aggregate losses there too. In Section 6, we use our empirical strategy to evaluate the effects of bank lending on a counterfactual policy change that would delete all bank defaults, which is similar to the informational change for non-banks after the policy change.

5. THE EFFECTS OF INFORMATION DELETION ON TOTAL SURPLUS

We present a simple framework adapted from Einav, Finkelstein, and Cullen (2010) and use our difference-in-difference estimates as inputs to the framework. Our focus is on understanding how deletion affects surplus and borrowing outcomes through adverse selection, not moral hazard. This is consistent with the empirical application we study here, a one-time deletion based on characteristics that were predetermined at the time of policy announcement.

[19] Credit cards are subject to a rate cap that was likely binding for retailer cards during this period.

Consider a consumer credit market where lenders set interest rates on the basis of observable borrower characteristics, but borrowers have private information on the cost of lending. Assume for simplicity that the lending market is competitive, so that in equilibrium rates are equal to average costs. As in Einav, Finkelstein, and Cullen (2010), lenders set rates, and quantities are endogenously determined.

Individual borrowers are denoted by i. Lenders partition markets using two types of borrower characteristics. The first type, X_i, is always observable to lenders. For the rest of this section, we think of the analysis as taking place within subgroups of borrowers defined by $X_i = x$. This captures the fact that in general, lenders offer different prices to observably different borrowers. The second type, $Z_i \in \{0, 1\}$, is a variable that will be deleted from the lender's information set, e.g. by the policy change. We model $Z_i = 1$ as being a default flag that predicts higher costs.[20]

Figure 7 shows the analysis graphically, with technical details available in Online Appendix B. The left panel describes the high-cost market ($Z_i = 1$) and the right panel describes the low-cost market ($Z_i = 0$). Because of adverse selection, marginal cost curves are downward sloping and equilibrium price and quantity in each market are determined by the intersection of market-specific *average* cost and demand curves. These are labeled, respectively, AC_{z_j} and D_{z_j}. q^e_j is the pre-deletion equilibrium quantity borrowed in market j.

The surplus-maximizing quantity and price in each market are in turn given by the intersection of market-specific demand and marginal cost curves, the latter labeled MC_{z_j}. Below we show evidence consistent with adverse selection in both markets, therefore of surplus losses, due to asymmetric information in both markets ex ante. In Figure 7, these losses are given by the

[20] To guarantee unique equilibria, we assume that the (inverse) demand curve crosses the marginal cost curve from above exactly once in both the high- and low-cost markets. For analytic tractability, we further assume that the demand and cost curves are linear.

areas of triangles A and B in in the high- and low-cost markets, respectively.

After deletion, lenders no longer observe Z_i and must set one price for both $Z_i = 0$ and $Z_i = 1$. The demand curve in the pooled market is given by the sum of market-specific demand curves, while the pooled average cost curve is a quantity-weighted sum of the market-specific average cost curves. Equilibrium price (AC^p) and quantities in the pooled market are determined by the intersection of the pooled AC curve and the pooled demand curve. The quantity borrowed in each market $Z_i = j$ (q_j^p) is given by the intersection of the market-specific demand curve and AC^p. Because of downward sloping average cost curves, borrowing rises (and prices fall) in the high-cost market and the reverse takes place in the low-cost market.

Changes in total surplus from pooling are determined by the relationship between the group-specific demand and cost curves and the pooled average costs. For individuals with $Z_i = 0$ at baseline, rising rates due to pooling increase surplus losses due to underprovision of credit (denoted by triangle D in the right panel of Figure 7). For individuals with $Z_i = 1$, the effects of pooling on surplus are ambiguous. If AC^p is above the point where the marginal cost and demand curves cross, the effects of the policy on surplus within this market are unambiguously positive, as pooling reduces the underprovision of credit due to adverse selection. If AC^p is below the efficient price, as in Figure 7, then the effects are unclear. Losses from underprovision in the segregated market may outweigh losses from overprovision in the pooled market (equal to the area of triangle C in the left panel of Figure 7). As we discuss in more detail in Online Appendix B, we can obtain analytic solutions for these quantities given observations of a) the unpooled quantities and costs, and b) slopes of the demand and cost curves in each market.

Figure 7: Equilibria for high- and low-cost markets and under pooling

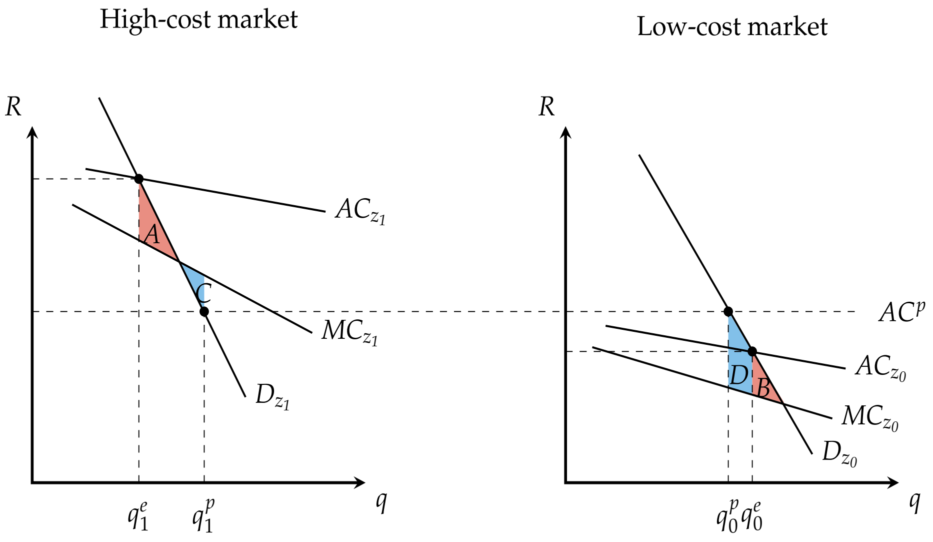


Diagram illustrating the economic framework. Left panel describes the high-cost market; right panel describes the low-cost market.

In general, the slopes of the demand and cost curves can be estimated using any exogenous shock to rates in each market. To tie our welfare analysis to the policy evaluation, we exploit shocks to lenders' predictions about borrowers' probability of default due to information deletion and use the results from the difference in differences analysis to estimate elasticities. We assume that the expected probability of default approximates bank's expectations of the cost of lending to an individual. Thus, under a policy of average cost pricing these shocks translate directly into rates. We map the high-cost and low-cost markets in the framework to the markets that face a reduction and an increase in predicted defaults in our empirical implementation, i.e., the markets with negative and positive exposure, respectively.

We estimate the slope of the demand curve in each market using results from Table II. To estimate the slope of the average cost curve, we use our difference-in-differences procedure to estimate the effect of deletion on *realized* costs in the high- and low-cost markets. We focus on a simple measure of realized costs: an indicator variable equal to one if a borrower adds to his default balance in the six month period following each registry snapshot. This is consistent with our assumption that defaults approximate lender costs. We estimate realized cost effects within the sample of individuals who increase their borrowing over the six-month period to recover cost curve slopes for market participants.

Online Appendix Tables 5 and 6 report the effects of deletion on realized average costs in the low-cost (top panel) and high-cost markets (low panel). Deletion slightly raises average costs for borrowers in the low-cost group and lowers average costs in the high-cost group. Because quantities fall in the low-cost group and rise in the high-cost group, the signs of these point estimates are consistent with downward-sloping average cost curves, and thus with adverse selection, in both markets. However, in neither case can we reject an effect of zero at conventional levels of significance. [21]

5.1. *Benchmark estimates of the effect of deletion on surplus*

As a benchmark we consider a market with no mark-up above average costs.[22] The level of $AC(x, z)$ is 0.029 (43% lower than average) and 0.069 (36% higher than average), while the average quantity borrowed is 252 and 113 thousand pesos in the low- and high-cost markets, respectively. Average cost curves slope down in both markets, leading to underprovision relative

[21] In Online Appendix Table 7 we repeat the analysis from Tables 5 and 6 using one-year-ahead bank default rather than six-month-ahead bank default to proxy for costs. Estimated effects of deletion on borrowing levels are close to unchanged relative to the benchmark analysis. We prefer our benchmark estimates because using one-year-ahead default measures means that some defaults attributed to loans originated in the pre-deletion period occur following deletion, which does not occur when we use the six-month-ahead measure.

[22] Online Appendix Figure 5 show the empirical demand, average cost, and marginal cost curves in the benchmark low-cost, high-cost, and pooled markets. Online Appendix Table 10 summarizes the quantitative implications of this analysis.

to the efficient quantity. Demand is less elastic in the high-cost market than the low-cost market. In our linear parameterization, the share of high-cost types in the market is equal to one for $R > 0.14$.

The equilibrium rate and average quantity in the pooled market, given by the intersection of the pooled demand curve and the pooled average cost curve, are $(q, R) = (215, 0.035)$. In the low-cost market, quantity borrowed declines by an average of $13,000 CLP per person, or a total of $26.4 billion CLP while The surplus loss relative to the efficient quantity rises by 106% of the baseline value. In contrast, rates in the high-cost market drop from 0.069 to 0.035, and borrowing rises by $28,000 CLP per person, or $17 billion CLP in aggregate. Welfare losses in this market decline by 73%. Aggregating across markets, borrowing falls by $9 billion CLP, and surplus losses rise by an amount equal to 66% relative to baseline.[23]

5.2. *Markups over average cost*

If borrowers face imperfect competition and are able to mark up prices relative to our cost measures, our benchmark analysis will systematically underestimate how much consumers value borrowing.[24] Further, if borrowers in the high- and low-cost markets face *different* markups at baseline, we will mismeasure their relative valuations. To explore how different assumptions about markups in the high- and low-cost markets affect our analysis, we add a market-specific markup term m_j for rates relative to average costs, so that for each market j, $R^e_j = (1 + m_j) \times AC^e_j$. In the pooled market, we allow a markup of value m_p over average costs. Within this framework we fix the low-cost market markup m_0 at a value μ_0, and set the high-cost market

[23] In Online Appendix Table 8 we re-do the analysis using one-year ahead default as a proxy for cost, rather than six-months ahead. In aggregate, surplus losses are larger in levels but smaller in percentage terms (42%) due to larger estimates of welfare losses at baseline.

[24] Ausubel (1991) shows evidence of lack of competition in the US credit card market.

markup m_1 to $m_1 = \mu_0 \times (1 + \mu_1)$. We cycle through combinations of μ_0 and μ_1, in each case setting m_p to the quantity-weighted average markup in the pre-deletion period so that deletion does not affect the average markup in the market.[25]

Figure 8 show the percentage changes in surplus loss relative to baseline value in both markets combined for different combinations of μ_0 and μ_1.[26] Surplus losses persist as we raise markups in both markets equally. As markups rise, both losses in the low-cost market and gains in the high-cost market rise in absolute value. This makes sense: higher markups mean that the consumers in both markets place a higher value on borrowing, leading to higher welfare stakes. Net losses rise in levels but fall in percentage terms due to a larger denominator.

Augmenting the markup in the high-cost market relative to the low-cost baseline tends to reduce the surplus losses from pooling. Again, this makes sense. Higher markups for high-cost borrowers mean that those individuals value borrowing more. At baseline markup levels up to 25%, surplus losses persist for additional high-cost markups of up to 100%. The effects of pooling on total surplus become zero or modestly positive in percentage term when markups are very high overall, *and* there are large additional markups in the high-cost market. According to our analysis, the deletion policy breaks even in surplus terms when, a) overall markups are large, and b) markups in the high cost market are larger relative to the low cost market. For example, we find that pooling breaks even in surplus terms when the low-cost markup is 50% and the additional high-cost markup is 100%,

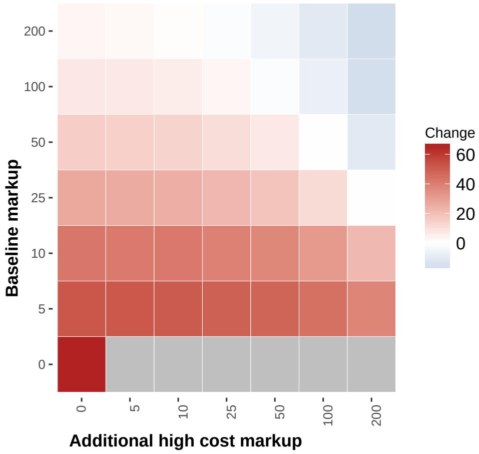
[25] The assumption that pooling does not affect the average markup may be violated if deletion affects market power (Mahoney and Weyl, 2017). However, Figure 1 shows that after the deletion the median consumer credit interest rate increases 20% by (5.3 percentage points from a base of 26%), which is similar to to the estimated 22% increase in predicted default for the low-cost market (the median borrower is not in default, i.e. low cost), shown in Table II, column 3 (i.e. rates and defaults increase *proportionally* following deletion).

[26] Online Appendix Table 9 presents the results.

and may even reduce surplus losses relative to the efficient outcome by 11% when the low-cost market markup is 200% and the high-cost market markup is an additional 100%.

We also note that deletion may have dynamic welfare effects (Handel, Hendel, and Whinston (2015), Clifford and Shoag (2016), Bartik and Nelson (2016), Cortes, Glover, and Tasci (2016), and Kovbasyuk and Spagnolo (2018)) or welfare effects outside of the credit markets (Bos, Breza, and Liberman, 2018; Herkenhoff, Phillips, and Cohen-Cole, 2016; Dobbie, Goldsmith-Pinkham, Mahoney, and Song, 2016). One can view our findings as measures of the costs of providing these benefits.

Figure 8: Heatmap of % change in welfare loss relative to baseline loss



Percent changes in total surplus loss relative to baseline loss reported in Table 7 under different assumptions about markups in low- and high-cost markets. Surplus calculations described in section 5. Vertical axis is markup at baseline in both high- and low-cost markets. Horizontal axis is additional markup in high-cost market. Average markups are constant before and after deletion.

6. EVALUATION OF COUNTERFACTUAL DELETION POLICIES

The methodology used above to study the effects of the large-scale deletion of credit bureau defaults provides a framework through which policymakers can predict the distributional and aggregate effects of changes in any type of credit information. In this section we apply this methodology to two hypothetical changes in the credit information available to lenders. The first is a deletion of information about gender. The idea of eliminating the use of demographic information has parallels in US anti-discrimination laws as applied to credit markets (Munnell, Tootell, Browne, and McEneaney, 1996; Blanchflower, Levine, and Zimmerman, 2003; Pope and Sydnor, 2011). The second is deletion of banks' internal and external default records across all banks in addition to the credit bureau defaults. This is a more radical version of the original policy.

In each case, we can simulate the effects of counterfactual policies using the following procedure. First, we compute each individual's (log) exposure to the policy by estimating predicted costs with and without the deleted information. We then take our estimates of exposure to cost changes, and scale them by an estimated elasticity of borrowing with respect to costs. For example, we can use the elasticity estimates from Table II.

We present the analysis in Table 11 for our baseline analysis. For each of the two counterfactual policies, we split the sample into individuals whose costs increase by 15% or more, individuals whose costs decrease by 15% or more, and the zero change group, which groups everyone else. This follows the procedure from our analysis of the observed deletion policy.

The top panel presents the first counterfactual policy, deletion of the gender indicator. Three things emerge from the analysis. First, most individuals (87% of the sample) belong to the zero change group. This is because the distribution of changes in costs is much tighter than in our baseline analysis, as is evident in the histogram of exposures shown in Figure 4. Second, as expected, gender is a strong predictor of cost changes: 98% of individuals exposed to cost increases are female, while females only represent 16% of those exposed to cost decreases. Thus, women would experience average increases in predicted costs following a deletion of the gender flag. Third, individuals whose costs increase or decrease have no registry defaults, and little variation in socio-economic status. These variables have little explanatory power for changes in banks' expected costs following deletion of the gender flag, which is consistent with the fact that costs do not change much when gender is deleted.

The bottom panel shows the second counterfactual policy, deletion of banks' internal default records in addition to consolidate default. Unsurprisingly, the more radical deletion option leads to larger changes in predicted costs than the actual deletion policy, as only 13% of the distribution is concentrated in the zero change group. This point is also shown in Figure 4. This suggests that the measure of defaults is highly predictive of future bank costs. Second, gender is uncorrelated with changes in costs following deletion of bank defaults. While bank defaults are, unsurprisingly, highly correlated with changes in predicted costs. Finally, socio-economic status is also correlated with changes in predicted costs: individuals exposed to reductions in costs are about 20 percent more likely to belong to a low socio-economic status group than those exposed to increases.

If one is willing to assume that elasticities of borrowing with respect to changes in average costs are the same as what we observe in the analysis of the observed deletion policy, we can go beyond the analysis of changes in the predicted cost distribution and predict the effects of these counterfactual deletion policies on borrowing. For example, if we take an estimated elasticity of -0.29 from Table 1 and multiply by the mean measures of exposure to the gender deletion in each group, we get that groups exposed to increases in costs see a 7 percent decline in new borrowing, a decline of $4,400 CLP per borrower, while groups exposed to decreases in costs see a 7.3 percent increase in new borrowing, an increase of $5,600 CLP per borrower. Multiplying each effect by the number of individuals in each group implies a near-zero change in aggregate new borrowing. The counterfactual deletion of banks' default records leads to a 18% drop in lending for individuals exposed to increases in costs and a 25% increase in lending for individuals exposed to decreases in costs. These effects aggregate to a drop in lending of \$42 billion CLP over a six month period, roughly twice the size of the \$20 billion CLP net effect of the observed deletion policy.

7. CONCLUSION

This paper explores the equilibrium effects of information asymmetries on credit markets in the context of a large-scale policy change that forced credit bureaus to stop reporting past defaults for the majority of defaulters in the Chilean consumer credit market.

We document a large increase in consumer credit rates as the information deletion policy is implemented and information on default is no longer reported. To quantify how the policy affects the information available to lenders we use a machine learning alorithm to summarize how the policy affected the ability to predict default amoung consumers. We find that some populations are affected significantly, while the information available from other sources leads others to not be affected directly. To estimate the causal effects of deletion on consumer credit borrowing, we implement a difference-in-differences test that compares the evolution of borrowing for individuals whose predicted bank default increases or decreases as a consequence of the deletion of information relative to individuals whose predicted bank default does not change. Our core empirical finding is that losses from information deletion are regressive and outweigh gains in this setting: consumer borrowing falls by 3.5% after the policy change, with the largest losses for lower-income individuals with smaller borrowing balances. Using a simple framework, we estimate the effects of the policy change on total surplus under several assumptions of bank pricing policies. There is no evidence that the winners from the policy value borrowing sufficiently more than the losers to offset these losses.

Our findings suggest that policies that limit information availability in credit markets can produce significant negative effects through general equilibrium effects on aggregate lending and rates even though some individuals can benefit. A feature of deletion policies is that the biggest losers tend to resemble the biggest winners on all characteristics observable to the lender other than the deleted information, so policies implemented with the goal of helping disadvantaged populations also have a greater risk of negative effects for these populations.

Our findings motivate a simple procedure by which policymakers can predict the distributional consequences of a proposed change in credit information. The procedure is to construct default and cost predictions before and after the change and identify the individuals with the biggest gains and losses in predicted costs. These estimates can be used alone to classify likely winners and losers, can be paired with estimates of demand elasticities to predict changes in quantity borrowed, or can be combined with estimates of demand and cost elasticities to predict changes in surplus. This approach can also be applied to understanding how existing information-restricting institutions such as sunset provisions affect lending. We leave this exercise for future research.

REFERENCES

AGAN, A., AND S. STARR (2017): “Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment,” *The Quarterly Journal of Economics*, 133(1), 191–235.

AGARWAL, S., S. CHOMSISENGPHET, N. MAHONEY, AND J. STROEBEL (2018): “Do Banks Pass Through Credit Expansions to Consumers who Want to Borrow?,” *Quarterly Journal of Economics*, 133(1).

AKERLOF, G. A. (1970): “The Market for ”Lemons”: Quality Uncertainty and the Market Mechanism,” *The Quarterly Journal of Economics*, 84(3), 488–500.

ATHEY, S., AND G. IMBENS (2016): “Recursive Partitioning for Heterogeneous Causal Effects,” *Proceedings of the National Academy of Sciences*, 113(27), 7353–7360.

Athey, S., and S. Wagner (2017): “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests,” *Working Paper*.

Ausubel, L. M. (1991): “The Failure of Competition in the Credit Card Market,” *The American Economic Review*, 81(1), 50–81.

Bartik, A. W., and S. Nelson (2016): “Credit Reports as Resumes: The Incidence of Pre-Employment Credit Screening,” *Working Paper*.

Bester, H. (1985): “Screening vs. Rationing in Credit Markets with Imperfect Information,” *American Economic Review*, 75(4), 850–55.

Blanchflower, D. G., P. B. Levine, and D. J. Zimmerman (2003): “Discrimination in the small-business credit market,” *The Review of Economics and Statistics*, 85(4), 930–943.

Bos, M., E. Breza, and A. Liberman (2018): “The Labor Market Effects of Credit Market Information,” *Review of Financial Studies*, 31(6), 2005–2037.

Bos, M., and L. I. Nakamura (2014): “Should Defaults be Forgotten? Evidence from Variation in Removal of Negative Consumer Credit Information,” Discussion paper, FRB of Philadelphia Working Paper.

Breiman, L. (2001): “Random Forests,” *Machine Learning*, 45, 5–32.

Breiman, L., J. Friedman, C. J. Stone, and R. Olshen (1984): *Classification and Regression Trees*. Chapman and Hall/CRC.

Brown, M., and C. Zehnder (2007): “Credit Reporting, Relationship Banking, and Loan Repayment,” *Journal of Money, Credit and Banking*, 39(8), 1883–1918.

Burlig, F., C. Knittel, D. Rapson, M. Reguant, and C. Wolfram (2017): “Machine Learning From Schools About Energy Efficiency,” *NBER Working Paper*, (w23908).

Clifford, R., and D. Shoag (2016): “”No More Credit Score” Employer Credit Check Banks and Signal Substitution,” *Working Paper*.

Cortes, K., A. Glover, and M. Tasci (2016): “The Unintended Consequences of Employer Credit Check Bans on Labor and Credit Markets,” *Working Paper*.

Cowan, K., and J. De Gregorio (2003): “Credit Information and Market Performance: The Case of Chile,” in *Credit Reporting Systems and the International Economy*, ed. by M. J. Miller, vol. 4, pp. 163–201. MIT Press, Cambridge, MA.

Dobbie, W., P. Goldsmith-Pinkham, N. Mahoney, and J. Song (2016): “Bad Credit, No Problem? Credit and Labor Market Consequences of Bad Credit Reports,” Discussion Paper 22711, National Bureau of Economic Research.

Dobbie, W., A. Liberman, D. Paravisini, and V. Pathania (2018): “Measuring

Bias in Consumer Lending," Working Paper 24953, National Bureau of Economic Research.

Einav, L., A. Finkelstein, and M. R. Cullen (2010): "Estimating Welfare in Insurance Markets Using Variation in Prices," *The Quarterly Journal of Economics*, 125(3), 877–921.

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

Elul, R., and P. Gottardi (2015): "Bankruptcy: Is It Enough to Forgive or Must We Also Forget?," *American Economic Journal: Microeconomics*, 7(4), 294–338.

Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther (2017): "Predictably Unequal? The Effects of Machine Learning on Credit Markets," Discussion paper, National Bureau of Economic Research.

González-Uribe, J., and D. Osorio (2014): "Information Sharing and Credit Outcomes: Evidence from a Natural Experiment," Discussion paper, Working Paper.

Handel, B., I. Hendel, and M. D. Whinston (2015): "Equilibria in health exchanges: Adverse selection versus reclassification risk," *Econometrica*, 83(4), 1261–1313.

Herkenhoff, K., G. Phillips, and E. Cohen-Cole (2016): "The impact of consumer credit access on employment, earnings and entrepreneurship," Discussion paper, National Bureau of Economic Research.

Huang, C.-L., M.-C. Chen, and C.-J. Wang (2007): "Credit Scoring with a Data Mining Approach Based on Support Vector Machines," *Expert Systems with Applications*, 33(4), 847–856.

Jaffee, D. M., and T. Russell (1976): "Imperfect Information, Uncertainty, and Credit Rationing," *The Quarterly Journal of Economics*, pp. 651–666.

Khandani, A. E., A. J. Kim, and A. W. Lo (2010): "Consumer Credit-Risk Models via Machine-Learning Algorithms," *Journal of Banking & Finance*, 34(4), 2767–2787.

Kovbasyuk, S., and G. Spagnolo (2018): "Memory and markets," Discussion paper, Working Paper.

Kulkarni, S., S. Truffa, and G. Iberti (2018): "Removing the Fine Print: Standardization, Disclosure, and Consumer Loan Outcomes," Discussion paper, Working Paper.

Liberman, A. (2016): "The Value of a Good Credit Reputation: Evidence from Credit Card Renegotiations," *Journal of Financial Economics*, 120(3), 644–660.

Mahoney, N., and E. G. Weyl (2017): "Imperfect competition in selection markets," *Review of Economics and Statistics*, 99(4), 637–651.

Miller, M. J. (2003): *Credit reporting systems and the international economy.* Mit Press.

Mullainathan, S., and J. Spiess (2017): "Machine Learning: An Applied EconometricAapproach," *Journal of Economic Perspectives*, 31(2), 87–106.

Munnell, A. H., G. M. Tootell, L. E. Browne, and J. McEneaney (1996): "Mortgage lending in Boston: Interpreting HMDA data," *The American Economic Review*, pp. 25–53.

Musto, D. K. (2004): "What Happens when Information Leaves a Market? Evidence from Postbankruptcy Consumers," *The Journal of Business*, 77(4), 725–748.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011): "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, 12, 2825–2830.

Petersen, M. A., and R. G. Rajan (2002): "Does Distance Still Matter? The Information Revolution in Small Business Lending," *The Journal of Finance*, 57(6), 2533–2570.

Pope, D. G., and J. R. Sydnor (2011): "What's in a Picture? Evidence of Discrimination from Prosper. com," *Journal of Human Resources*, 46(1), 53–92.

Rothschild, M., and J. E. Stiglitz (1976): "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information," *The Quarterly Journal of Economics*, 90(4), 630–49.

Steinberg, J. (2014): "Your privacy is now at risk from search engines– even if the law says otherwise," *Forbes*.

Stiglitz, J., and A. Weiss (1981): "Credit Rationing in Markets with Imperfect Information," *The American Economic Review*, 71(3), 393–410.

Varian, H. (2016): "Causal Inference in Economics and Marketing," *Proceedings of the Natural Academy of Sciences*, 113(27), 7310–7315.

1. ONLINE APPENDIX - ADDITIONAL TABLES AND RESULTS

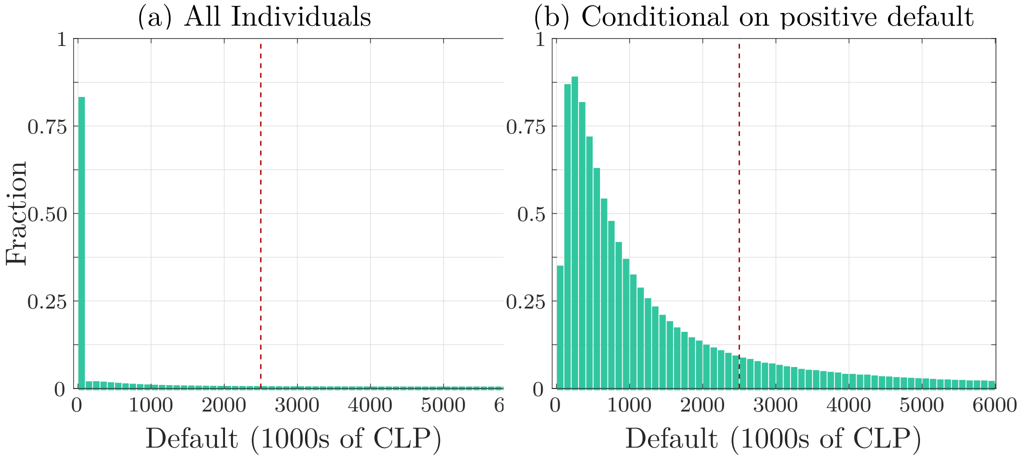
TABLE 1

Demographics by exposure category

	Positive exposure	Zero group	Negative exposure	Pooled
Positive Default	0.01	0.46	0.99	0.31
Amt. Default	52	696	456	566
New Borrowing	236	175	99	195
New Debt	468	356	156	384
Positive Bank Default	0.04	0.15	0.18	0.10
Low SES	0.50	0.56	0.71	0.55
Have Mortgage	0.25	0.18	0.18	0.22
Age	44.4	43.8	42.5	43.9
Female	0.47	0.41	0.46	0.45
Share of individuals	0.53	0.32	0.16	1
N	2,051,138	1,234,733	612,737	3,898,608

Baseline borrowing and demographic characteristics by exposure-generated market type in July 2011. Rows correspond to features of the sample and columns define market type. 'Positive default' is an indicator for whether individuals have positive default balances within the snapshot while 'Amt. Default' computes the mean default value conditional on having positive default. 'New borrowing' computes mean new borrowing across all individuals, as does new 'New debt.' 'Positive bank default' indicates positives bank default for individuals within the snapshot. 'Low SES' is an indicator flagging bank defined socioeconomic status.' Have mortgage' is an indicator flagging whether individuals have positive mortgage balances in the snapshot. 'Age' reports the mean age of individuals in the snapshot in years. 'Female' is flags gender reported to the bank. Share of individuals computes the share of total individuals in the snapshot contained in each market, while *N* reports the number of individuals (observations).

Figure 1: Histogram of ammount in default as of December 2011

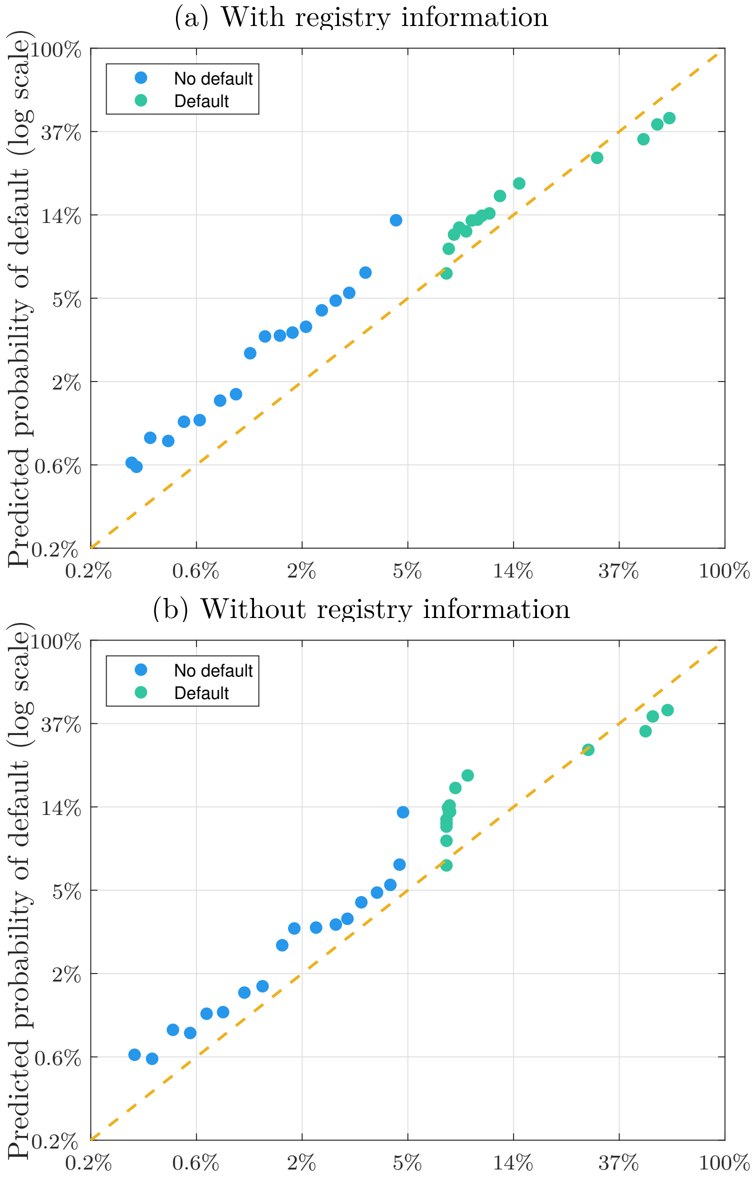


(a) All Individuals

(b) Conditional on positive default

Note: The left panel shows a histogram of consolidated defaults as of December 2011, for amounts below \$6 million CLP (approximately \$3,000). The right panel shows a histogram of consolidated defaults for individuals with positive defaults only.

Figure 2: Predictions with and without registry data



(a) With registry information

(b) Without registry information

Binned means of random forest default predictions (horizontal axis; log scale) vs. out-of-sample observed default outcome (vertical axis, log scale). Top panel uses predictions that include registry information. The bottom panel uses predictions that exclude registry information. Our default outcome measure is an indicator variable for at least one new default in the six month period beginning in February 2012, the date of registry deletion. Predictions are constructed using registry and borrowing data from December 2009 and June 2010. See text for details.

TABLE 2

DiD by exposure, mortgage, and socioeconomic status

	Positive Exposure			
	Predicted Defaults		New Borrowing	
By Mortgage Status				
	No Mortgage	Mortgage	No Mortgage	Mortgage
Jun. 2010	0.03	−0.05*	−5.21	−3.84
	(0.04)	(0.02)	(3.31)	(4.22)
Dec. 2010	0.02	−0.05^{+}	1.04	4.48
	(0.04)	(0.03)	(3.29)	(5.66)
Dec. 2011	0.20***	0.22***	−13.22***	−8.85
	(0.04)	(0.04)	(3.72)	(6.91)
Elasticity			−0.35	−0.13
Dep. Var. Base Period Mean	0.05	0.03	185.39	318.06
N Clusters	303	292	303	293
N Obs.	2,204,290	706,443	10,148,532	2,945,193
N Individuals	1,432,239	437,433	3,566,538	923,617
N Exposed Individuals	375,676	129,619	1,609,450	522,605
By Socioeconomic Status				
	Low SES	High SES	Low SES	High SES
Jun. 2010	0.04	−0.00	−0.40	−2.78
	(0.05)	(0.02)	(3.59)	(3.59)
Dec. 2010	0.02	−0.02	1.61	−1.59
	(0.04)	(0.02)	(3.09)	(4.22)
Dec. 2011	0.22***	0.21***	−8.78***	−21.31***
	(0.05)	(0.03)	(2.58)	(4.82)
Elasticity			−0.41	−0.30
Dep. Var. Base Period Mean	0.07	0.02	95.12	347.84
N Clusters	303	302	303	302
N Obs.	1,147,411	1,763,322	6,999,869	6,093,856
N Individuals	849,835	1,064,389	2,768,287	2,021,242
N Exposed Individuals	216,450	288,845	1,109,738	1,022,317

Significance: $^{+}$ 0.10 * 0.05 ** 0.01 *** 0.001.Difference in difference estimates from equation 2 over defined subsamples. Columns 1 through 4 are predicted default and borrowing diff-in-diff effect estimates in the high exposure market while columns 5 through 8 report estimates in the low exposure market. Column headers report dependent variable at the top and subsample below. Sample in specifications where default is an outcome conditions on positive borrowing (see text for details). 'Elasticity' is borrowing effect scaled by base period outcome mean and predicted default effect within each market-subsample. We take the log of 'Predicted Default' for estimation but report the base period mean in levels. 'N exposed individuals reports the number of individuals not in the 0 group included in the regression sample in the treatment period. Since some individuals appear in multiple snapshots we report both individuals and observations. Standard errors clustered at market level. See text for details.

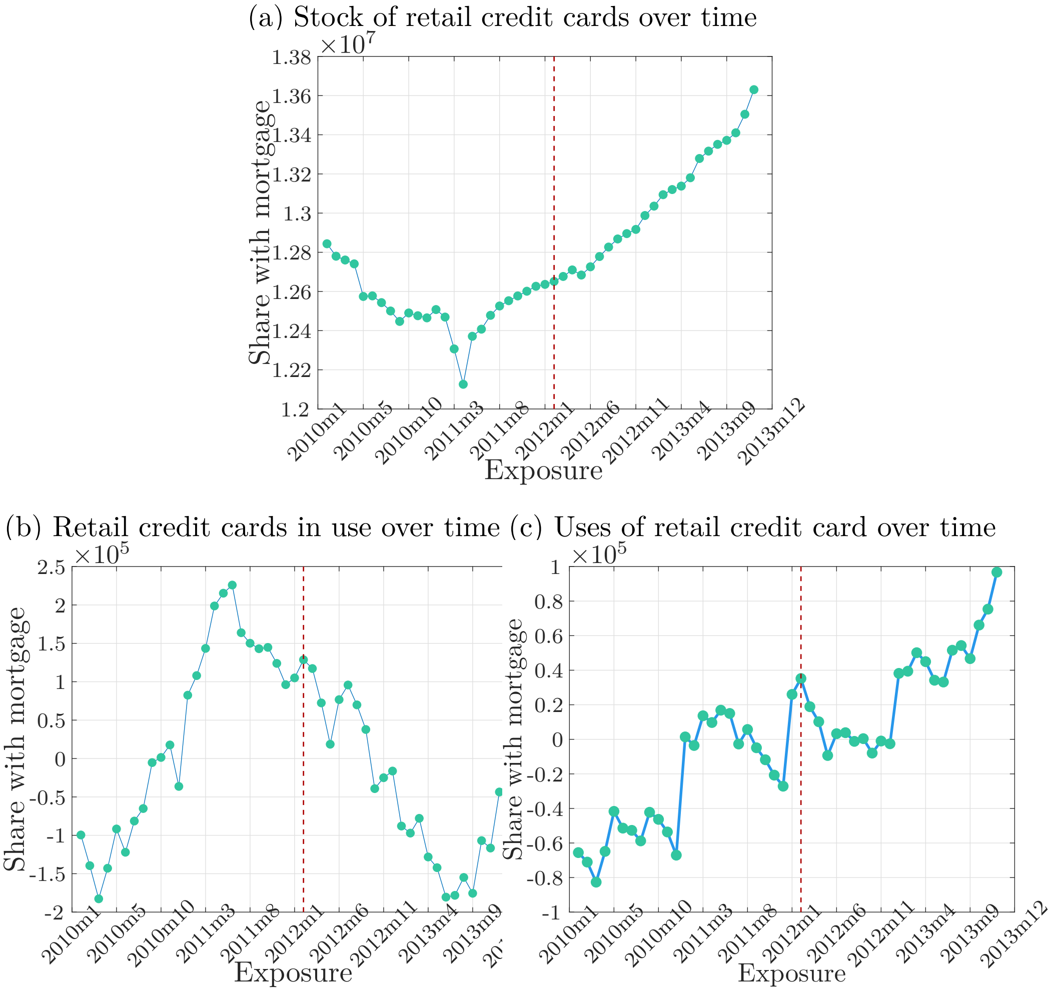
TABLE 3

DiD by exposure, mortgage, and socioeconomic status

	Negative Exposure			
	Predicted Defaults		New Borrowing	
By Mortgage Status				
	No Mortgage	Mortgage	No Mortgage	Mortgage
Jun. 2010	0.11	-0.10	-6.08^{*}	-0.78
	(0.08)	(0.06)	(2.56)	(4.33)
Dec. 2010	0.07	-0.09^{+}	0.46	5.59
	(0.08)	(0.05)	(2.81)	(4.16)
Dec. 2011	-0.27^{***}	-0.42^{***}	15.73^{***}	19.78^{***}
	(0.07)	(0.05)	(4.06)	(5.11)
Elasticity			-0.46	-0.23
Dep. Var. Base Period Mean	0.1	0.09	127.19	204.06
N Clusters	278	266	281	272
N Obs.	1,028,499	244,872	6,135,611	1,358,357
N Individuals	800,061	193,751	2,649,628	606,131
N Exposed Individuals	70,162	14,584	497,783	110,446
By Socioeconomic Status				
	Low SES	High SES	Low SES	High SES
Jun. 2010	0.12	-0.04	-1.32	-6.32
	(0.09)	(0.05)	(2.89)	(4.15)
Dec. 2010	0.08	-0.05	-1.03	6.47
	(0.08)	(0.04)	(2.55)	(4.53)
Dec. 2011	-0.30^{***}	-0.32^{***}	9.27^{**}	18.78^{***}
	(0.07)	(0.05)	(3.05)	(5.47)
Elasticity			-0.41	-0.24
Dep. Var. Base Period Mean	0.16	0.05	75.44	243.48
N Clusters	274	282	279	285
N Obs.	555,634	717,737	4,617,114	2,876,854
N Individuals	471,664	532,229	2,021,269	1,378,643
N Exposed Individuals	56,279	28,467	421,652	186,577

Significance: $^{+}$ 0.10 * 0.05 ** 0.01 *** 0.001.Difference in difference estimates from equation 2 over defined subsamples. Columns 1 through 4 are predicted default and borrowing diff-in-diff effect estimates in the high exposure market while columns 5 through 8 report estimates in the low exposure market. Column headers report dependent variable at the top and subsample below. Sample in specifications where default is an outcome conditions on positive borrowing (see text for details). 'Elasticity' is borrowing effect scaled by base period outcome mean and predicted default effect within each market-subsample. We take the log of 'Predicted Default' for estimation but report the base period mean in levels. 'N exposed individuals reports the number of individuals not in the 0 group included in the regression sample in the treatment period. Since some individuals appear in multiple snapshots we report both individuals and observations. Standard errors clustered at market level. See text for details.

Figure 3: Retail credit cards over time



(a) Stock of retail credit cards over time

(b) Retail credit cards in use over time

(c) Uses of retail credit card over time

Panel (a) Stock of retail credit cards by month. Time of deletion policy noted with vertical line; Panel (b) Number of retail credit cards used by month. Time of deletion policy noted with vertical line. Source: SBIF; Panel (c) Amount of retail credit purchases by month. Time deletion policy noted with vertical line. Source: SBIF.

TABLE 4

Log Likelihoods of Various Algorithms

	Pre-period		Contemporaneous	
	Training	Testing	Training	Testing
Naive Bayes				
With registry info	−0.412	−0.682	−0.398	−0.633
Without registry info	−0.324	−0.516	−0.300	−0.458
Logistic LASSO				
With registry info	−0.176	−0.324	−0.176	−0.335
Without registry info	−0.180	−0.337	−0.182	−0.348
Random Forest				
With registry info	−0.176	−0.278	−0.173	−0.295
Without registry info	−0.180	−0.284	−0.177	−0.305

Mean binomial log likelihoods for each algorithm. Columns identify the sample in which the log likelihood value is calculated. The ‘training’ sample is a 10% random sample of borrowers with new borrowing in the July 2009 Snapshot (pre-period) and within each snapshot (contemporaneous). ‘Testing’ identifies the main sample used in our analysis, from which the training set is dropped. Rows identify prediction methods. Within each prediction method, the ‘with registy info’ row uses registry information in addition to the other, while the ‘without registry info’ row does not. See section 4 for the full list of predictors and Appendix A for details on the transformation of these predictors and the structure of each algorithm.

TABLE 5

Difference in difference estimates on realized default

Positive Exposure			
	Pooled	No Mortgage	Have Mortgage
Jun. 2010	0.02	0.03	-0.05^{*}
	(0.03)	(0.04)	(0.02)
Dec. 2010	0.00	0.01	-0.05^{+}
	(0.03)	(0.04)	(0.03)
Dec. 2011	0.02	0.01	0.03
	(0.03)	(0.04)	(0.03)
Elasticity	0.10	0.05	0.12
Dep. Var. Base Period Mean	0.04	0.05	0.03
N Clusters	303	303	292
N Obs	4,930,411	3,734,294	1,196,117
N Individuals	2,385,366	1,894,374	558,811
N Exposed Individuals	855,928	636,066	219,862

Negative Exposure			
	Pooled	No Mortgage	Have Mortgage
Jun. 2010	0.07	0.11	−0.10
	(0.08)	(0.08)	(0.06)
Dec. 2010	0.04	0.06	-0.09^{+}
	(0.07)	(0.08)	(0.05)
Dec. 2011	−0.04	−0.03	−0.08
	(0.06)	(0.07)	(0.05)
Elasticity	0.12	0.11	0.18
Dep. Var. Mean	0.1	0.1	0.09
N Clusters	284	281	268
N Obs	2,156,891	1,742,719	414,172
N Individuals	1,433,629	1,167,470	284,204
N Exposed Individuals	143,165	118,441	24,724

Significance: $^{+}$ 0.10 * 0.05 ** 0.01 *** 0.001. Difference and difference estimates from equation 2 where the dependent variable is realized default 'Registry information' reports the estimated effect of deletion while other column headers only define subsamples. Columns 2-6 are estimated over the low cost market (as defined by exposure) while columns 7-11 are over the high cost market. We take the log of 'Realized default' when estimating the regressions but report the base period mean in levels. 'Elasticity' is the realized default effect scaled by the predicted default effect. 'N exposed individuals reports the number of individuals not in the 0 group included in the regression sample in the treatment period. Since some individuals appear in multiple snapshots we report both individuals and observations. Standard errors clustered at market level. See text for details.

TABLE 6

Difference in difference estimates on realized default

Positive Exposure		
	Low SES	High SES
Jun. 2010	0.04	−0.00
	(0.05)	(0.02)
Dec. 2010	0.01	−0.02
	(0.04)	(0.02)
Dec. 2011	0.01	0.03
	(0.04)	(0.03)
Elasticity	0.06	0.16
Dep. Var. Base Period Mean	0.07	0.02
N Clusters	303	302
N Obs	1,943,879	2,986,532
N Individuals	1,201,766	1,347,265
N Exposed Individuals	366,368	489,560

Negative Exposure		
	Low SES	High SES
Jun. 2010	0.12	−0.04
	(0.09)	(0.05)
Dec. 2010	0.07	−0.05
	(0.08)	(0.04)
Dec. 2011	−0.06	−0.02
	(0.07)	(0.05)
Elasticity	0.20	0.05
Dep. Var. Mean	0.16	0.05
N Clusters	278	283
N Obs	941,603	1,215,288
N Individuals	721,292	755,356
N Exposed Individuals	94,855	48,310

Significance: $^{+}$ 0.10 * 0.05 ** 0.01 *** 0.001. Difference and difference estimates from equation 2 where the dependent variable is realized default 'Registry information' reports the estimated effect of deletion while other column headers only define subsamples. Columns 2-6 are estimated over the low cost market (as defined by exposure) while columns 7-11 are over the high cost market. We take the log of 'Realized default' when estimating the regressions but report the base period mean in levels. 'Elasticity' is the realized default effect scaled by the predicted default effect. 'N exposed individuals reports the number of individuals not in the 0 group included in the regression sample in the treatment period. Since some individuals appear in multiple snapshots we report both individuals and observations. Standard errors clustered at market level. See text for details.

TABLE 7

Difference-in-difference predictions using long run default measures

	Positive exposure			*Negative exposure*		
	Predicted Default	Average Cost	New Borrowing	Predicted Default	Average Cost	New Borrowing
Jun. 2010	0.01	0.00	−7.09*	0.03	0.03	−5.68$^+$
	(0.02)	(0.02)	(3.05)	(0.05)	(0.05)	(3.23)
Dec. 2010	0.01	0.01	−2.11	0.02	0.01	0.30
	(0.02)	(0.02)	(3.52)	(0.05)	(0.05)	(3.25)
Jun. 2011	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Dec. 2011	0.25***	0.12***	−13.28**	−0.30***	0.04	17.98***
	(0.02)	(0.02)	(4.21)	(0.04)	(0.04)	(3.47)
Elasticity		0.48	−0.24		−0.12	−0.36
Dep.Var. Mean	0.08	0.08	214.70	0.14	0.14	165.09
N Clusters	307	307	307	299	299	300
N Obs.	2,929,133	4,961,674	13,163,613	1,486,567	2,519,339	8,117,207
N Individuals	1,844,615	2,394,399	4,373,700	1,104,246	1,571,258	3,422,263
N Exposed Ind.	452,132	765,941	1,967,865	79,572	134,306	589,628

Significance: $^+$ 0.10 * 0.05 ** 0.01 *** 0.001. Difference and difference estimates from equation 2. Table is identical to Table II but uses a one-year ahead measure of default to compute predicted default rates. See section 4 for details. The first two columns report the difference-in-difference estimated effect of deletion on outcome variables listed in column headers, while the third and fourth estimate the dif-in-dif effect on the different exposure-defined markets. We take the log of 'Predicted default' for estimation but report the base period mean in levels. 'Elasticity' is borrowing effect scaled by base period outcome mean and predicted default effect. '*N* exposed individuals' reports the number of individuals not in the zero group included in the regression sample in the treatment period. Since some individuals appear in multiple snapshots we report both individuals and observations. Standard errors clustered at market level.

TABLE 8

Distribution of deletion effects using long run default measures

	Separate	Pooled	Difference
Positive exposure			
Predicted cost	0.065	0.081	0.016
Average cost	0.065	0.073	0.008
New borrowing (1000s CLP)	234.779	222.246	−12.533
Surplus loss (1000s CLP)	1.711	2.138	0.427
Aggregate new borrowing (Bns CLP)	447	424	−24
Aggregate surplus loss (1000s CLP)	3,261,672	4,075,579	813,908
			24.95%
N individuals	1,905,946	1,905,946	1,905,946
Negative exposure			
Predicted cost	0.120	0.081	−0.039
Average cost	0.120	0.125	0.005
New borrowing (1000s CLP)	112.490	132.079	19.589
Surplus loss (1000s CLP)	0.140	1.128	0.988
Aggregate new borrowing (Bns CLP)	67	78	12
Aggregate surplus loss (1000s CLP)	83,086	668,656	585,570
			704.77%
N individuals	592,732	592,732	592,732
Combined			
Average cost	0.072	0.081	0.008
New borrowing (1000s CLP)	205.770	200.857	−4.913
Surplus loss (1000s CLP)	1.339	1.899	0.560
			41.84%
Aggregate new borrowing (Bns CLP)	514	502	−12
Aggregate surplus loss (1000s CLP)	3,344,758	4,744,236	1,399,478
			41.84%
N individuals	2,498,678	2,498,678	2,498,678

This table describes changes in key metrics before and following deletion, with inputs to the theoretical framework using the long-run cost measure, assuming a 0% markup.

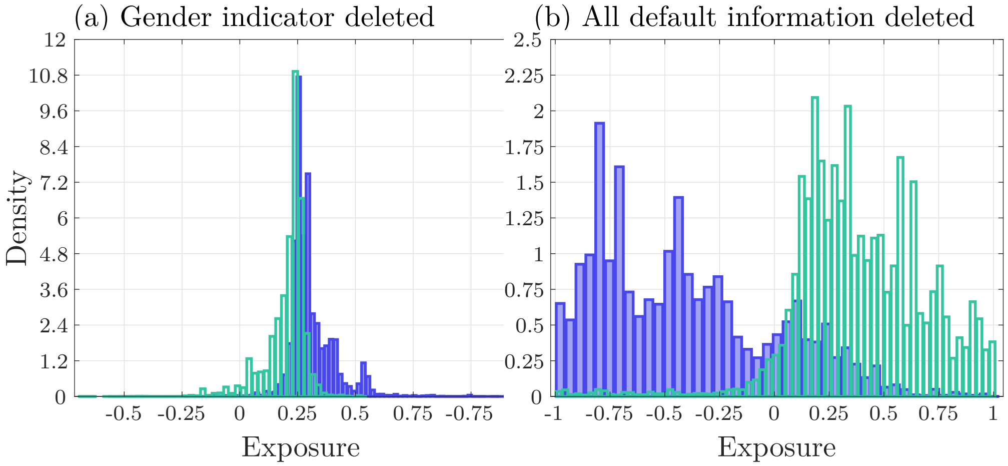
TABLE 9

Surplus changes by markup

		Additional high cost market markup (%)					
		0	5	10	25	50	100
Low cost market markup (%)	0	0.17 −0.11 0.10 65.52%					
	5	0.19 −0.19 0.10 51.94%	0.19 −0.19 0.10 51.40%	0.19 −0.19 0.10 50.87%	0.19 −0.20 0.10 49.27%	0.19 −0.22 0.10 47.95%	0.20 −0.25 0.09 44.09%
	10	0.21 −0.26 0.10 42.35%	0.21 −0.27 0.10 41.47%	0.21 −0.27 0.10 41.77%	0.21 −0.29 0.10 39.16%	0.22 −0.33 0.09 37.19%	0.23 −0.40 0.08 31.24%
	25	0.27 −0.48 0.10 26.58%	0.27 −0.49 0.10 26.05%	0.27 −0.51 0.09 24.66%	0.28 −0.57 0.09 22.28%	0.30 −0.66 0.08 18.25%	0.32 −0.86 0.05 11.01%
	50	0.37 −0.84 0.09 15.01%	0.38 −0.88 0.09 14.52%	0.38 −0.92 0.08 13.43%	0.40 −1.04 0.07 10.28%	0.43 −1.25 0.04 6.02%	0.49 −1.68 −0.01 −0.72%
	100	0.57 −1.56 0.08 7.23%	0.59 −1.65 0.07 6.46%	0.60 −1.74 0.06 5.33%	0.64 −2.00 0.03 2.49%	0.71 −2.46 −0.02 −1.59%	0.85 −3.38 −0.12 −7.79%
	200	0.98 −3.00 0.06 2.81%	1.01 −3.20 0.04 1.85%	1.03 −3.39 0.02 0.71%	1.13 −3.97 −0.04 −1.81%	1.28 −4.94 −0.15 −5.57%	1.61 −6.88 −0.35 −11.18%

This table describes changes in changes in surplus loss before and following deletion. Cells are additional markups (columns, in percent terms) relative to a given markup rate in the low cost market (rows). Within each cell, rows are level changes in surplus loss in the low cost, high cost, mean change in surplus loss across both markets, and percent change in surplus loss relative to baseline loss the pooled market following deletion.

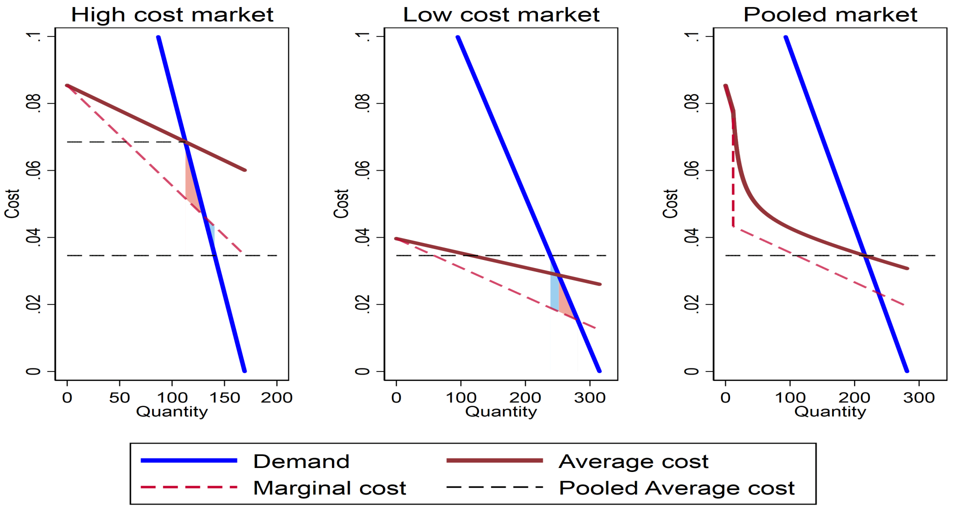
Figure 4: Distribution of exposure under counterfactual deletion policies


(a) Gender indicator deleted

(b) All default information deleted

Histograms of exposure under counterfactual deletion policies. On top: log difference in prediced defaults ('exposure') excluding and including a gender indicator variable, split by gender. Below: exposure defined when all default information is deleted from the credit registry, split by default amout. See text for details. Predictions trained within each month.

Figure 5: Empirical estimates of different markets



Empirical estimate of figure 7 using difference-in-difference estimates of slopes, assuming average cost pricing in both markets. See Section 4 for details.

TABLE 10

Distribution of deletion effects

	Separate	Pooled	Difference
Positive exposure			
Predicted cost	0.029	0.035	0.006
Average cost	0.029	0.029	0.001
New borrowing (1000s CLP)	251.561	238.714	−12.847
Surplus loss (1000s CLP)	0.161	0.331	0.170
Aggregate new borrowing (Bns CLP)	516	490	−26
Aggregate surplus loss (1000s CLP)	330,480	679,717	349,238
			105.68%
N individuals	2,051,138	2,051,138	2,051,138
Negative exposure			
Predicted cost	0.069	0.035	−0.034
Average cost	0.069	0.064	−0.004
New borrowing (1000s CLP)	112.713	140.695	27.981
Surplus loss (1000s CLP)	0.156	0.041	−0.114
Aggregate new borrowing (Bns CLP)	69	86	17
Aggregate surplus loss (1000s CLP)	95,456	25,307	−70,149
			−73.49%
N individuals	612,737	612,737	612,737
Combined			
Average cost	0.033	0.035	0.001
New borrowing (1000s CLP)	219.624	216.168	−3.455
Surplus loss (1000s CLP)	0.160	0.265	0.105
			65.52%
Aggregate new borrowing (Bns CLP)	585	576	−9
Aggregate surplus loss (1000s CLP)	425,936	705,025	279,089
			65.52%
N individuals	2,663,875	2,663,875	2,663,875

This table describes changes in key metrics before and following deletion. Prices and surplus calculations assume average cost pricing. See text for details. 'Positive exposure' panel is individuals whose predicted defaults rise following deletion; 'Negative exposure' is individuals whose predicted defaults fall. 'Combined' panel averages over both markets for prices, average cost, new borrowing, and surplus measures, while summing for aggregate borrowing/surplus measures. 'New borrowing' in 1000s of CLP. Aggregate new borrowing is in billions of CLP.

TABLE 11

Effects of counterfactual exposure policies

	Exposed to pred. default increases	Zero group	Exposed to pred. default decreases	Pooled
Gender deleted				
Exposure increases	0.24	0.00	-0.25	0.00
Positive Default	0.00	0.34	0.00	0.36
Amt. Default	479	571	71	1,621
New Borrowing	63	184	81	168
New Debt	203	369	106	337
Positive Bank Default	0.02	0.10	0.04	0.10
Low SES	0.18	0.22	0.17	0.22
Have Mortgage	0.08	0.21	0.12	0.20
Age	45.3	43.9	45.5	44.1
Female	0.98	0.44	0.16	0.45
Share of individuals	0.04	0.87	0.04	1
N	171,878	4,111,244	166,565	4,721,885
All info. deleted				
Exposure increases	0.63	0.06	-0.84	0.15
Positive Default	0.07	0.18	0.93	0.36
Amt. Default	460	432	602	1,621
New Borrowing	135	535	77	168
New Debt	307	985	128	337
Positive Bank Default	0.06	0.08	0.20	0.10
Low SES	0.22	0.16	0.26	0.22
Have Mortgage	0.22	0.18	0.18	0.20
Age	44.4	45.2	42.5	44.1
Female	0.46	0.43	0.44	0.45
Share of individuals	0.55	0.13	0.25	1
N	2,615,689	630,130	1,203,868	4,721,885

Baseline borrowing and demographic characteristics by exposure-generated market type in July 2011 under counterfactual policy changes. Panels are separated by counterfactual policy: deleting a gender indicator variable and deleting all default information. Rows correspond to features of the sample and columns define market type. ‘Positive default’ is an indicator for whether individuals have positive default balances within the snapshot while ‘Amt. Default’ computes the mean default value conditional on having positive default. ‘New borrowing’ computes mean new borrowing across all individuals, as does new ‘New debt.’ ‘Positive bank default’ indicates positives bank default for individuals within the snapshot. ‘Low SES’ is an indicator flagging bank defined socioeconomic status.‘ Have mortgage’ is an indicator flagging whether individuals have positive mortgage balances in the snapshot. ‘Age’ reports the mean age of individuals in the snapshot in years. ‘Female’ is flags gender reported to the bank. Share of individuals computes the share of total individuals in the snapshot contained in each market, while *N* reports the number of individuals (observations).

A. ONLINE SUPPLEMENT A - ADDITONAL PREDICTION MODEL DETAILS

We generate cost predictions by regressing an indicator for new default against a large selection of features using a random forest algorithm. We create four sets of predictions trained on 10% of the data with new borrowing within each snapshot – approximately 8% of the overall data. Predictions are trained and predicted either contemporaneously, within each 6-month post-December snapshot (PD^{post}), or only in the December 2009 snapshot (PD^{pre}). The random forests for each type are constructed with or without registry information. We use `python`'s `sklearn` package to perform our machine learning tasks (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot, and Duchesnay, 2011).

Our random forest regression design constructs regression trees using a feature vector of the following observable characteristics of each observation: a gender indicator, and one and two period lags of innovations in borrowing, innovations in total debt, total borrowing, total debt, average costs, and credit line information. We additionally include the default history deleted from the credit registry in some of the trees. In total, these trees have either thirteen or fourteen predictor variables.

We scale our features by binning their nonzero values into quartiles. This reduces noise in the feature vector and creates parsimonious regression trees. In our dataset, we find that this additionally decreases the time necessary to construct a random forest. Finally, we subset over only new borrowers in each period so that our cost estimates reflect costs conditional on borrowing.

To generate our PD^{pre} predictions, we train a model only using observations in the December 2009 snapshot. PD^{post} predictions are generated using a training sample from each snapshot; these predictions are actually generated using a suite of models each tied to a particular snapshot.

We use three-fold cross validation combined with a grid search to pick parameters for each model. The parameters over which we search are the minimum number of observations in a terminal node (*minleaf*) and the number of features over which each tree can sample. We set the number of trees in a forest to 150. Predictive power is not sensitive to choices in this range. See figures A1 and A2 to see outcomes from this procedure.

Constructing random forests is (generally) a supervised learning task. Breiman (2001) defines a random forest as a set of regression trees, $h_k = h(x, \Theta_k)$ where h is a tree and Θ_k is a random selection of observations and features from the training data, where each tree "votes" on the output given an observation. We pick splits in the data to reduce mean-squared error, as is common with regression tasks. We use this loss function and a regression task, despite our target variable existing only in $\{0, 1\}$, to ensure that our outputs are continuous on $[0, 1]$ and reflect probabilities. Our predictions are best thought of as a weighted average of default rate in pools of observations clustered together by similarity along a set of their covariates.

We additionally estimate a regression tree[27] to bin borrowers into smaller markets. We define a market as a set of observations M such that $h(x_i, \Theta)$ returns a prediction stemming from the same terminal node for all $i \in M$. We use this method to cluster borrowers into borrowers with similar features and default rates. These clusters therefore represent infered groups in the data at the level which we believe the treatment is applied and are analagous to the clusters defined in each tree in the forest.

Finally, we recreate the analysis above, exchanging the random forest algorithm for two other machine learning procedures that return classification probabilities. These are a naive Bayes classifier and a logistic LASSO. Our naive Bayes classifier first bins nonzero values along the feature vector into quartiles. Under the naive assumption of independence of features in

[27] We estimate CART-style regression trees that split using variance reduction (Breiman, Friedman, Stone, and Olshen, 1984).

the feature vector, the classifier constructs $P(\text{default}|X)$ using Bayes' formula under the assumption that $P(X|\text{default})$ is Gaussian, though this is functionally irrelevant due to binning.

For the logistic LASSO, we take the log of nonzero values of continuous features, dummying out zero values using indicator variables. We perform a logistic regression with a λ penalty term of the sum absolute value of the coefficients and use three-fold cross validation to pick λ for each model.

Finally, we classify observations' socioeconomic status by training a random forest classifier on observations for whom the bank defined socioeconomic status group. Our three-fold cross validation procedure indicates that we are able to do this with approximately 35% accuracy using a random forest composed of 100 trees and built on a feature vector consisting of continuous measures of consumer debt, mortgage amount, debt balance, credit line, bank default, average cose, age, total default amount, and indicators for gender, new borrowing, and having positive borrowing cap. See figure A3 for cross-validation output.

Figure A1: Cross-validation output forPD^{pre} random forest predictions

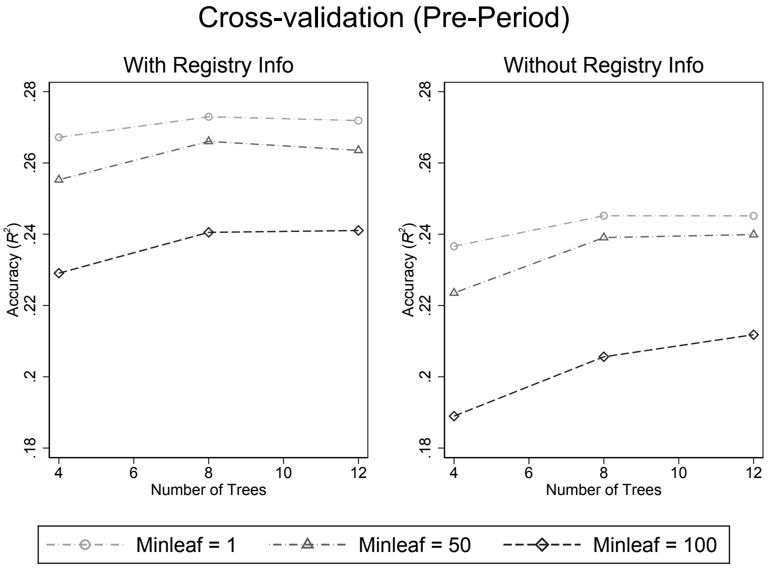


Figure A2: Cross-validation output for PD^{post} random forest predictions

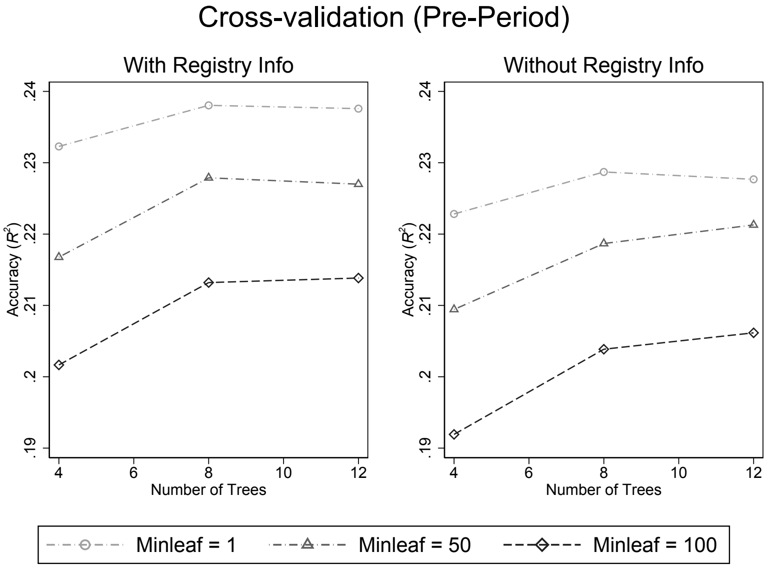
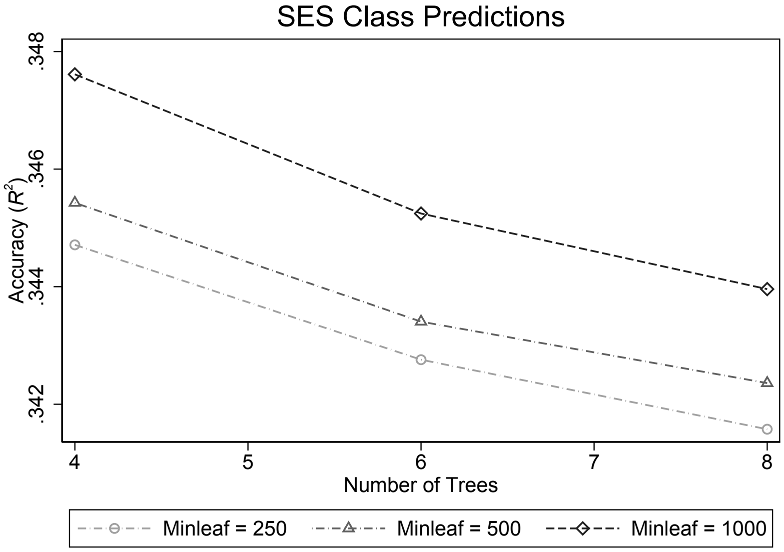


Figure A3: Cross-validation output for PD^{post} logistic LASSO predictions



B. ONLINE SUPPLEMENT B - ADDITONAL MODEL DETAILS

B.1. *Model setup*

This Appendix presents the details of main text Section 5. Model setup is as in the text. Let there be a unit measure of borrowers in the market, of whom a fraction α have $Z_i = 0$ and a fraction $1-\alpha$ have $Z_i = 1$. Demand and cost functions may vary across values of Z_i. Let $q_z(R)$, $MC_z(R)$, and $AC_z(R)$ denote the demand for credit, marginal cost, and average cost functions for type $Z_i = z$ as a function of the lender's (gross) offer rate R. $q_z(R)$ denotes the *average* quantity of credit purchased for individuals in the market, so that total market quantity is given by $\alpha q_0(R)$ for $Z_i = 0$ and $(1-\alpha)q_1(R)$ for $Z_i = 1$. To guarantee unique equilibria, we assume that the (inverse) demand curve crosses the marginal cost curve from above exactly once in each market. For analytic tractability, we further assume that the demand and cost curves are linear.

B.1.1. *Pre-deletion equilibria*

When lenders observe Z_i, equilbria are defined by the intersection of inverse demand and average cost curves in each market. Letting $R_z(q)$ represent the inverse demand curve in each market, equilibrium quantities q_z^e are determined by $R_z(q_z^e) = AC_z(R_z(q_z^e))$. Let $AC_z^e = AC_z(R_z(q_z^e))$ denote the equilibrium average cost in each market. We focus on the empirically relevant case where there is adverse selection in both markets; i.e., where marginal cost curves are downward sloping. The surplus-maximizing quantity q_z^* is determined by $R_z(q_z^*) = MC_z(q_z^*)$. We denote the surplus-maximizing rate as $R_z^* = R_z(q_z^*)$. Deadweight loss due to asymmetric information in market z is the area of the shaded triangle (denoted by "A" in the high cost market and "B" in the low-cost market in Figure 7, respectively), with total surplus loss in each market given by the formula:

$$(4) \qquad DWL_z = \frac{1}{2}\left(q_z^* - q_z(AC_z^e)\right) \times \left(AC_z^e - MC_z(AC_z^e)\right).$$

B.1.2. *Deletion policy*

In the pooling equilibrium lenders no longer observe Z_i. Demand in the pooled market at price R is given by $q(R) = q_0(R) + q_1(R)$, and the pooled market average cost is $AC(R) = s(R)AC_0(R) + (1 - s(R))AC_1(R)$, where the low-cost share $s(R)$ is defined as $s(R) = \frac{\alpha q_0(R)}{\alpha q_0(R)+(1-\alpha)q_1(R)},$. The equilibrium price/average cost AC^e and quantity q^e are determined by $AC^e = AC(R(q^e))$. The changes in average borrowing from pooling in each market are then given by:

$$\Delta q_z = q_z(AC^e) - q_z(AC_z^e),$$

and the average welfare loss by:

$$DWL_z = \frac{1}{2}\left(q_z^* - q_z(AC^e)\right) \times \left(AC^e - MC_z(AC^e)\right).$$

Changes in surplus from pooling are determined by the relationship between the group-specific demand and cost curves and the pooled average costs. For individuals with $Z_i = 0$ at baseline, rising rates due to pooling increase surplus losses due to underprovision of credit. These additional losses are denoted by D in the right panel of Figure 7, the low-cost market. For individuals with $Z_i = 1$, the effects of pooling on total surplus are ambiguous. If $AC^e > R_1^*$, then the effects of the policy for this group are unambiguously positive, as pooling reduces the underprovision of credit due to adverse selection. If $AC^e < R_1^*$, then the effects are unclear. Losses from overprovision in the pooled market may outweigh losses from underprovision in the segregated market. Figure 7 in the main text illustrates the latter case, with surplus losses from overprovision equal to the area of triangle C in the left panel.

B.1.3. *Measuring the effects of pooling*

The effects of pooling on equilibrium borrowing and surplus are determined by the slopes of the demand and cost curves in the high- and low-cost markets. Given observations of unpooled quantities q_z^e, costs AC_z^e, and slopes $\frac{dq_z}{dR}$ and $\frac{dAC_z}{dR}$, pooled equilibrium average costs and quantities are given by the solution to the system of equations

$$\begin{aligned} AC^p &= \frac{\alpha q_0^p}{\alpha q_0^p + (1-\alpha) q_1^p} AC_0^p + \frac{(1-\alpha) q_1^p}{\alpha q_0^p + (1-\alpha) q_1^p} AC_1^p \\ q_z^p &= q_z^e + \frac{dq_z}{dR}(AC^p - AC_z^e) \text{ for } z \in \{0,1\} \\ AC_z^p &= AC_z^e + \frac{dAC_z}{dR}(AC^p - AC_z^e) \text{ for } z \in \{0,1\} \end{aligned}$$

There are five equations and five unknowns, yielding an analytic solution for each value. Multiple equilibria are possible but, as we discuss below, not empirically relevant in the setting we consider here.

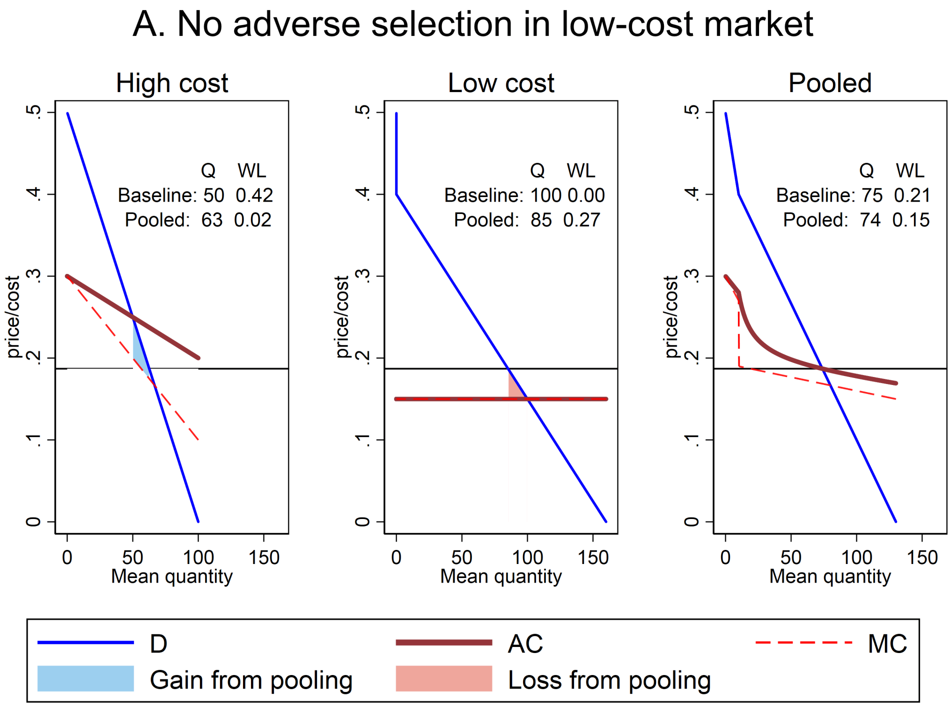
Computing effects on surplus requires knowledge of the levels and slopes of marginal cost curves in addition to the demand and average cost curves. Here we exploit the observation that the equilibrium value of marginal cost $MC_z^e = \frac{dAC_z}{dq} q_z^e + AC_z^e$, and that with linear average cost curves $\frac{dMC_z}{dq} = 2\frac{dAC_z}{dq}$.

In Appendix Figure B1, we simulate equilibrium outcomes from pooling a low-cost and high-cost market under different assumptions of the slopes of the demand and cost curves in each market. The figure illustrates how the effects of pooling on aggregate borrowing and total surplus are ambiguous, even in this simple model, and how they relate to the slopes of demand and costs.

B.1.4. *Alternative modeling approaches*

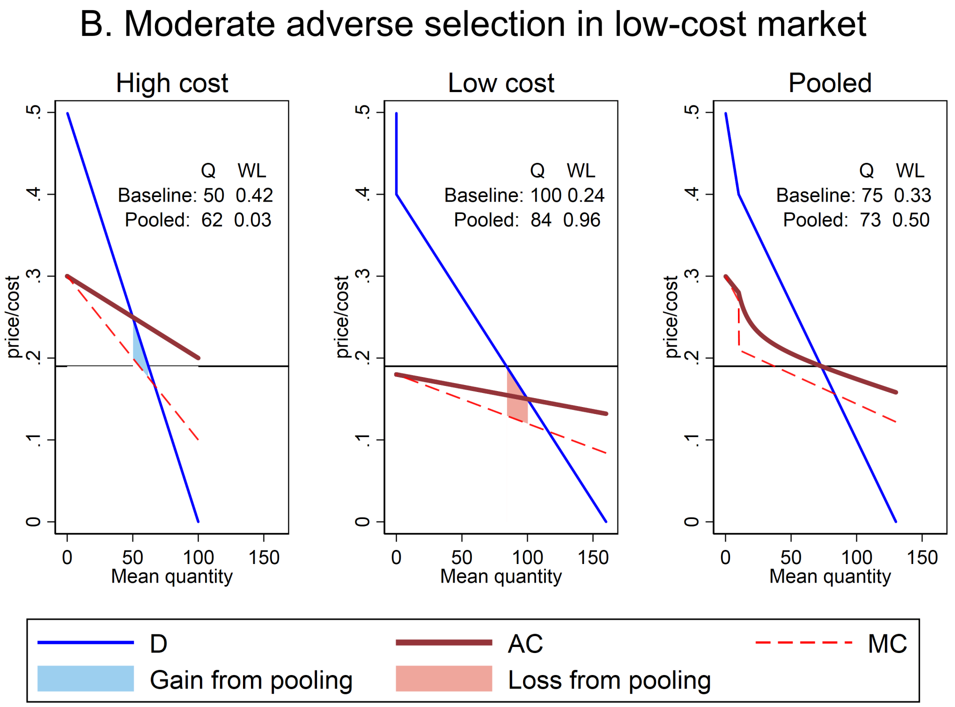
The framework we use to evaluate the consequences of the deletion policy on surplus here is one of several plausible modeling approaches. Most notably, we assume that lenders set prices rather than offering contracts consisting of rate-quantity pairs (Rothschild and Stiglitz, 1976), and that the form of contract does not change following policy implementation. This rules out separating equilibria where lenders screen borrowers based on their contract choice (Bester, 1985). In a simple screening model equilibrium, however, good types–non-defaulters– would have less credit than in the full information setting, while bad types–defaulters– would not have more credit. Because there is no counteracting positive effect for bad types, deletion increases surplus losses.

Figure B1: Simulated separating and pooling equilibria



Each panel shows simulated separate-market (left two panels) and pooled market (right panel) equilibria under different assumptions about market sizes and slopes of average cost and demand curves in the high-cost and low-cost market. Text in each panel displays aggregate market quantities transacted ("Q" column) and welfare loss relative to the efficient quantity ("WL" column) under the separate ("baseline") equilibrium and the "pooled" equlibrium. To see changes in aggregate welfare from pooling compare the "pooled" and "baseline" welfare loss columns in the rightmost panel of each row. Pooling causes welfare to rise in Panel A, fall in Panel B, and rise in Panel C. The number of individuals in high- and low-cost market are normalized to one. Separate equilibrium price and quantity (p, q) are the same in each panel, with $(p, q) = (0.25, 50)$ in the high market and $(p, q) = (0.15, 100)$ in the low-cost market. $\frac{dAC_0}{dq}$ and $\frac{AC_1}{dq}$ are the slopes of average cost curves in the low- and high-cost markets, respectively, with analogous definitions for demand curves. Slopes parameters vary across rows as follows. Panel A: $(0, -0.001, -400, -200)$ Panel B: $(-0.0003, -0.001, -400, -200)$. Panel C: $(0, -0.001, -400, -350)$. See text for model details.

Figure B2: (Cont'd) Simulated separating and pooling equilibria



B. Moderate adverse selection in low-cost market

High cost

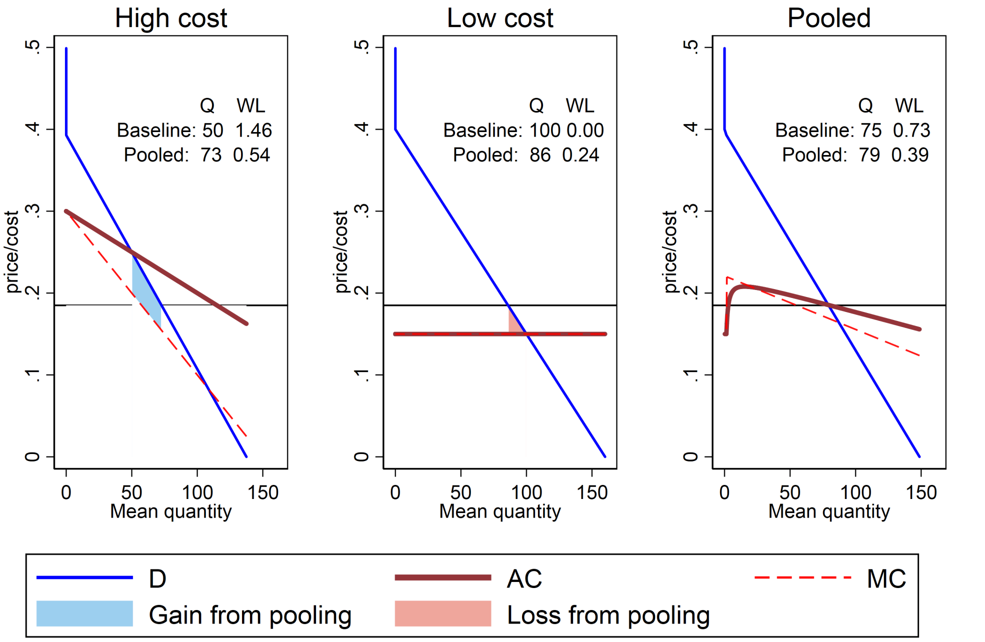
Low cost

Pooled

Each panel shows simulated separate-market (left two panels) and pooled market (right panel) equilibria under different assumptions about market sizes and slopes of average cost and demand curves in the high-cost and low-cost market. Text in each panel displays aggregate market quantities transacted ("Q" column) and welfare loss relative to the efficient quantity ("WL" column) under the separate ("baseline") equilibrium and the "pooled" equlibrium. To see changes in aggregate welfare from pooling compare the "pooled" and "baseline" welfare loss columns in the rightmost panel of each row. Pooling causes welfare to rise in Panel A, fall in Panel B, and rise in Panel C. The number of individuals in high- and low-cost market are normalized to one. Separate equilibrium price and quantity (p, q) are the same in each panel, with $(p, q) = (0.25, 50)$ in the high market and $(p, q) = (0.15, 100)$ in the low-cost market. $\frac{dAC_0}{dq}$ and $\frac{AC_1}{dq}$ are the slopes of average cost curves in the low- and high-cost markets, respectively, with analogous definitions for demand curves. Slopes parameters vary across rows as follows. Panel A: $(0, -0.001, -400, -200)$ Panel B: $(-0.0003, -0.001, -400, -200)$. Panel C: $(0, -0.001, -400, -350)$. See text for model details.

Figure B3: (Cont'd) Simulated separating and pooling equilibria

C. No adverse selection in low-cost market, less elastic demand



Each panel shows simulated separate-market (left two panels) and pooled market (right panel) equilibria under different assumptions about market sizes and slopes of average cost and demand curves in the high-cost and low-cost market. Text in each panel displays aggregate market quantities transacted ("Q" column) and welfare loss relative to the efficient quantity ("WL" column) under the separate ("baseline") equilibrium and the "pooled" equlibrium. To see changes in aggregate welfare from pooling compare the "pooled" and "baseline" welfare loss columns in the rightmost panel of each row. Pooling causes welfare to rise in Panel A, fall in Panel B, and rise in Panel C. The number of individuals in high- and low-cost market are normalized to one. Separate equilibrium price and quantity (p, q) are the same in each panel, with $(p, q) = (0.25, 50)$ in the high market and $(p, q) = (0.15, 100)$ in the low-cost market. $\frac{dAC_0}{dq}$ and $\frac{AC_1}{dq}$ are the slopes of average cost curves in the low- and high-cost markets, respectively, with analogous definitions for demand curves. Slopes parameters vary across rows as follows. Panel A: $(0, -0.001, -400, -200)$ Panel B: $(-0.0003, -0.001, -400, -200)$. Panel C: $(0, -0.001, -400, -350)$. See text for model details.