

Predicting Consumer Protection Risks through Administrative Data

Daniel Putman, Postdoctoral Fellow

November 3, 2021

Center for Financial Inclusion

Financial Inclusion Week

ipa
INNOVATIONS FOR POVERTY ACTION

IPA's Consumer Protection Research Initiative

Innovations for Poverty Action (IPA) is a research and policy nonprofit that creates and shares evidence, while equipping decision-makers to use evidence to reduce poverty.

Two research methods:

- Data collection, analysis and market monitoring
- Impact evaluation testing of new consumer protection solutions

Four key risks:

- Fraud in digital channels
- Consumer redress and complaints handling
- Product information and consumer choice
- Over-indebtedness

Bangladesh

Kenya

Nigeria

Uganda

Background

Market Monitoring Projects

Countries:

1. Kenya: with the Competition Authority of Kenya[1]
2. Sierra Leone: with UNCDF & the Bank of Sierra Leone[2]

Goals:

1. Identify potential consumer protection concerns
2. Inform development of policies

Research activities:

- Phone surveys
- Mystery shopping
- Social media monitoring
- Complaints analysis

1 Putman, D., Mazer, R., & Blackmon, W. (2021). Report on the Competition Authority of Kenya Digital Credit Market Inquiry Competition Authority of Kenya and Innovations for Poverty Action.
2 Blackmon, W., Cuccaro, F., Holzinger, A., Mazer, R., Ngwabe, W., & Putman, D. (2021). From the Field to Policy Formulation—How Research is Informing Consumer Protection in Sierra Leone. Innovations for Poverty Action. https://www.poverty-action.org/blog/field-policy-formulation—how-research-informing-consumer-protection-sierra-leone

What is administrative data?

Any data that is collected and stored by organizations for operational as opposed to research purposes

As digital credit has grown, automation has meant digitized record keeping, better for market monitoring

1. Digital credit administrative data has become larger scale and better kept
2. Includes records like account information, loan disbursements, fees, repayments, rollovers, etc.

What is the problem we wanted to solve with this project?

Tracking (over)indebtedness with transaction level data

- Overindebtedness is tricky to define and measure[3]
- However, regulators still need to be informed about debt levels

What did we come up with?

Administrative data on transactions provides several options in monitoring debt levels:

1. Track **default** and **late repayment** to understand symptoms
2. With data from several providers, find consumer who **multiple borrow**
3. Build a measure of **average indebtedness** to track debt levels directly

3 Garz, S., Giné, X., Karlan, D., Mazer, R., Sanford, C., & Zinman, J. (2020). Consumer Protection for Financial Inclusion in Low and Middle Income Countries: Bridging Regulator and Academic Perspectives (Vol. 2507, Issue 1). https://doi.org/10.1146/annurev-financial-071020-012008.

What is the problem we wanted to solve with this project?

Tracking (over)indebtedness with transaction level data

- Overindebtedness is tricky to define and measure[3]
- However, regulators still need to be informed about debt levels

What did we come up with?

Administrative data on transactions provides several options in monitoring debt levels:

In collaboration with the CAK, IPA conducted an analysis on 1 & 2

1. Track **default** and **late repayment** to understand symptoms
2. With data from several providers, find consumer who **multiple borrow**
3. Build a measure of **average indebtedness** to track debt levels directly

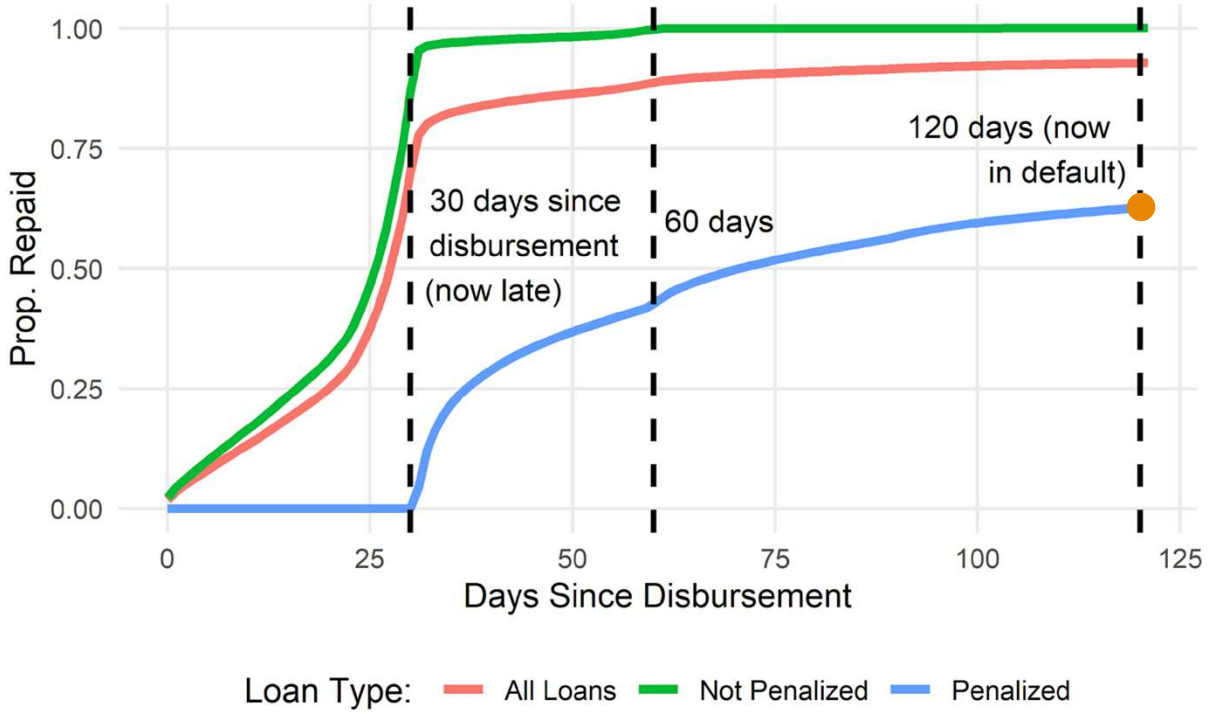
3 Garz, S., Giné, X., Karlan, D., Mazer, R., Sanford, C., & Zinman, J. (2020). Consumer Protection for Financial Inclusion in Low and Middle Income Countries: Bridging Regulator and Academic Perspectives (Vol. 2507, Issue 1). https://doi.org/10.1146/annurev-financial-071020-012008.

1. Track default and late repayment to understand symptoms

Administrative data let us look at rates of late repayment and default.

Indicators:

- Use **late fees** and **repayment behavior** to characterize if a borrower was late. Defined **late repayment** if borrower was charged a late fee
- Validated this with repayment behavior -- people who did not repay in 30 days but were not changed always repaid in 60
- Defined **default** if repaid in full 90 days after the due date and had been charged a late fee

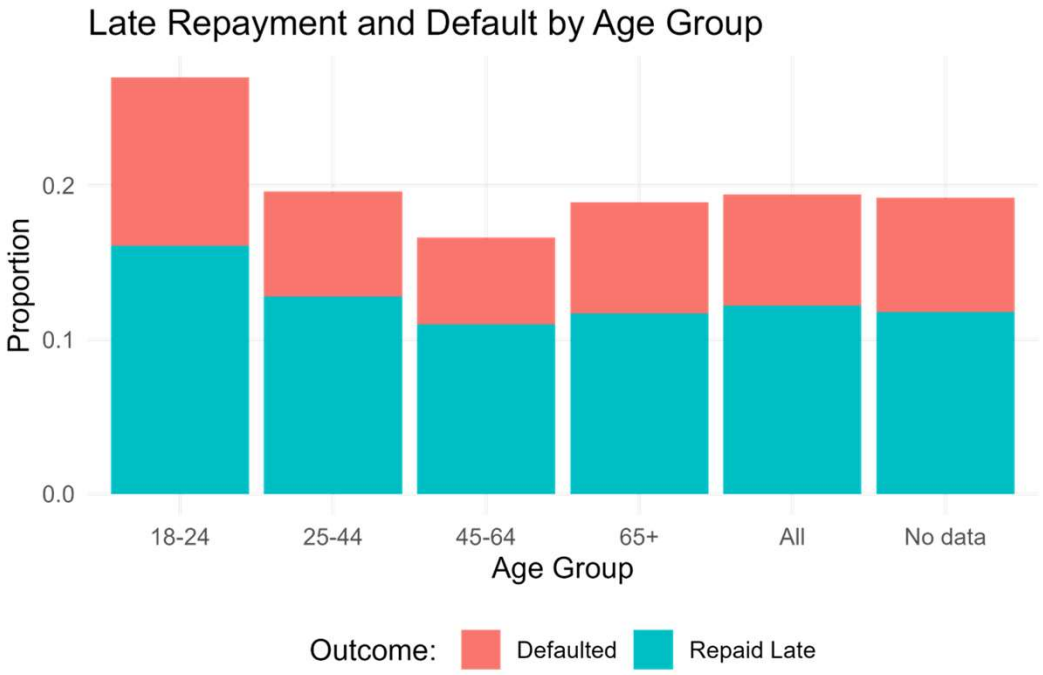


1. Track default and late repayment to understand symptoms

Which, in turn, provides insight into who these late borrowers are.

Insights from age disaggregation:

- Young borrowers are **habitually late and default more** than other groups.
- These borrowers may be high risk-high reward.

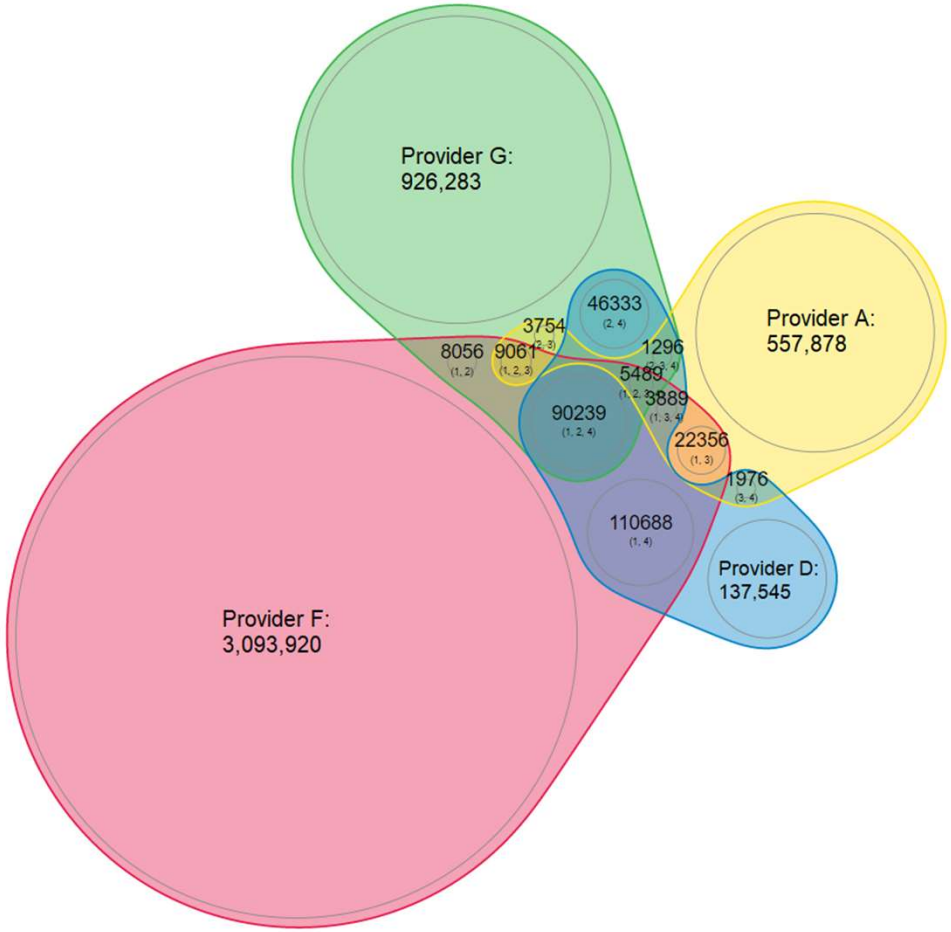


2. With data from several providers, find consumer who multiple borrow

Next, we identified individuals with multiple accounts.

Multiple Account Holding in the CAK Sample:

1. De-identified numbers through a common process which allowed us to match them later
2. Found 6% of our sample had multiple accounts



2. With data from several providers, find consumer who multiple borrow

Most individuals with multiple accounts have more than one credit accounts open at one time.

Multiple borrowing is borrowing from a different provider before the loan term at the first provider is done

82%

of people who multiple borrow within the sample of multiple account holders

Early/revolving borrowing is borrowing again from the same provider before the first loan term is done

87%

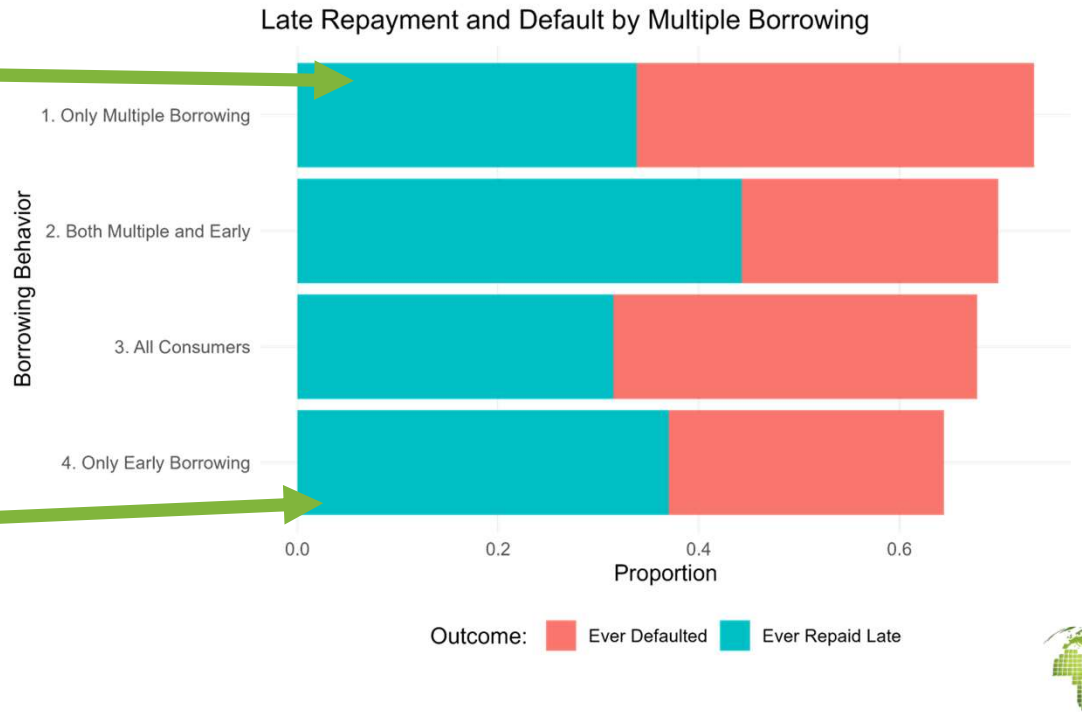
of people who early borrow within the sample of multiple account holders

2. With data from several providers, find consumer who multiple borrow

Merging the two datasets, we find that multiple borrowing from different providers is associated with higher rates of default.

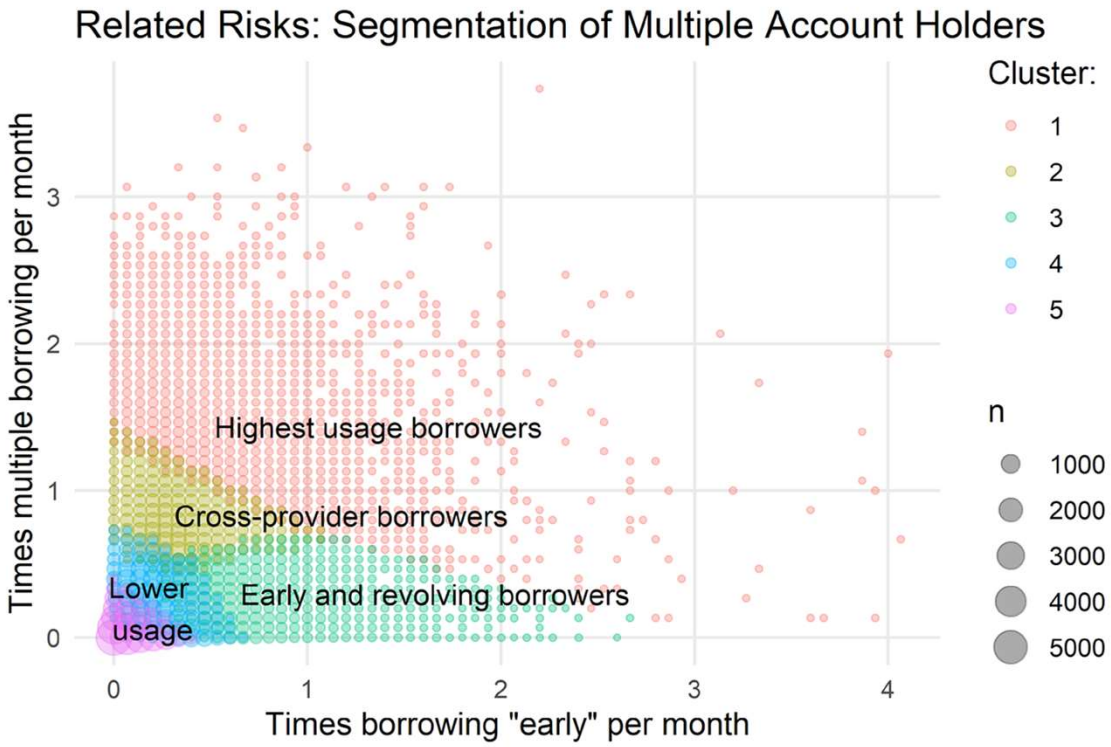
Those who *only* multiple borrow are late and default more than the average consumer

Those who only early borrow are late and default *less*



What else could we have done with the administrative data?

1. Look closer at **average indebtedness** across all accounts
2. Segmentation of borrowers
3. Predictive analysis:
   - Predict **default** using other measures
   - Link with survey data to predict other popular measures of over-indebtedness: **debt to income ratio**, **sacrifice-based over-indebtedness**,[4] **or financial well-being** outcomes [5]



4 Schicks, J. (2011). The over-indebtedness of microborrowers in Ghana-An empirical study from a customer protection perspective.
5 Consumer Financial Protection Bureau. (2015). Measuring financial well-being: A guide to using the CFPB Financial Well-Being Scale.

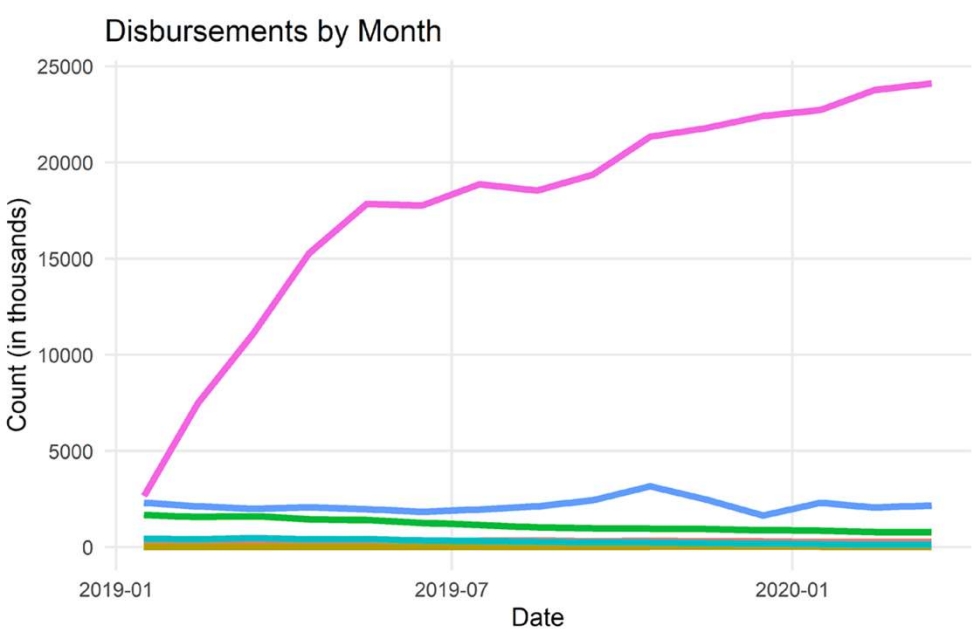
Should I Use Administrative Data?

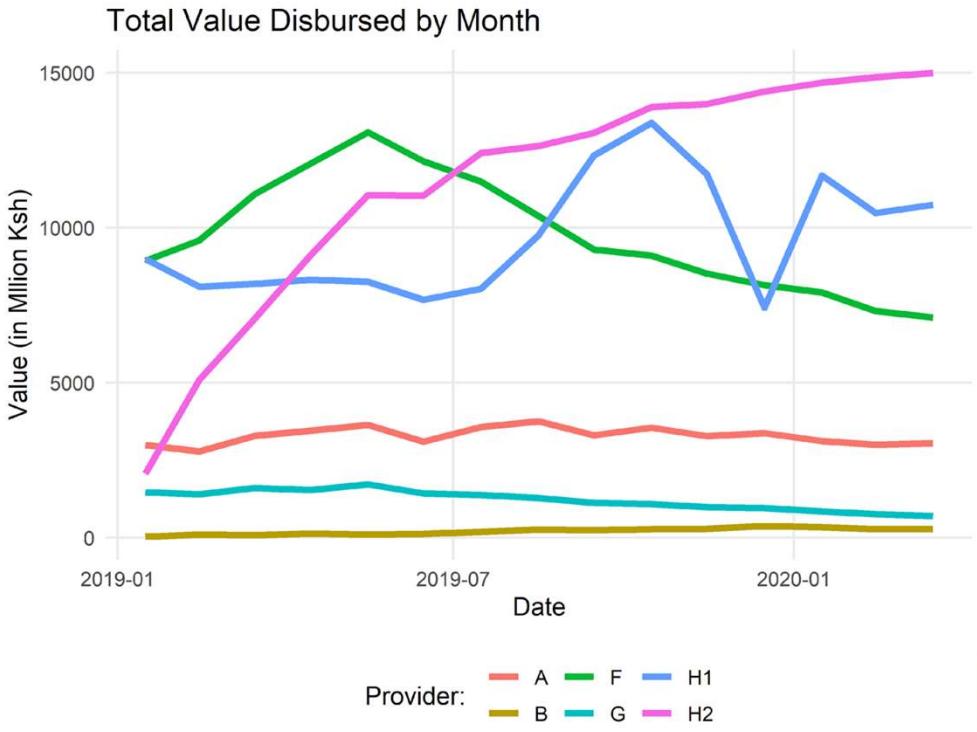
Advantages of administrative data

1. Useful in the very short term and the very long term
   - Detailed data for short term, up to date monitoring
   - Can trace out the evolution of outcomes over time
2. Better measurement of outcomes
   - Avoids biases from desirability
   - More accuracy and precision when studying difficult to recall outcomes
3. Lower cost for both participants and researchers

Example: Market dynamics in Kenya shifted significantly in 2019 with a new overdraft product

Product H2 became the largest lender by value by Q3 2019





Is administrative data the right tool for me?

1. Does my organization (or my partner organization) have the staffing to support obtaining and analyzing the data?
   - IT, data wrangling, analysis
2. Do I have the technical capacity and resources to manage the data request?
   - Talk to your IT folks!
   - Data storage, computing resources, etc.
3. Can I measure the relevant outcomes with administrative data?
   - Risks like high prices, defaults, debt, multiple borrowing
   - Evolution and concentration of the market

What can't administrative data measure? Some examples

1. Explicit misconduct by financial providers
   - Mystery shopping
   - Complaints
   - Social Media
2. Consumer knowledge and preferences about of digital credit
   - Surveys
   - Laboratory experiments

Anything else "external" to the record keeping

What data do I request?

Planning the data request by planning the analysis

Your guiding lights

- Pre-Analysis Plan
- IT System of providers
- Legal, ethical, and logistical constraints

What are you using the data for?

- Impact evaluation (RCT, Quasi-experimental)
  - *EX: regression discontinuity using data on entrance exam scores*
- Predictive analysis
  - *EX: predicting poverty using call detail records data, see Blumenstock et al. (2015)*[6]
- Descriptive analysis/monitoring tools
  - *EX: Digital Credit Market Inquiry with Competition Authority of Kenya*

While pre-analysis plans are common for RCTs, valuable as a sanity check!

6 Blumenstock, J. E., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073–1076.

Pre-Analysis Plans: Defining Variables of Interest

An Example from Financial Transactions Data

- Price of a loan could mean many things
- Annualized Percentage Rate requires **cost**, **loan size**, and **tenure**:

$$APR = \left(\frac{Cost}{Loan\ Size}\right)\left(\frac{365\ days}{Tenure}\right) \times 100\%$$

  - Cost and loan size are transaction level
  - However, tenure is loan level
- Therefore, we may need to request a loan level dataset that can be merged with transactions, would need **loan id** to link datasets
- Alternatively, if we aren't using transaction detail, we could ask for data summarized at the loan level

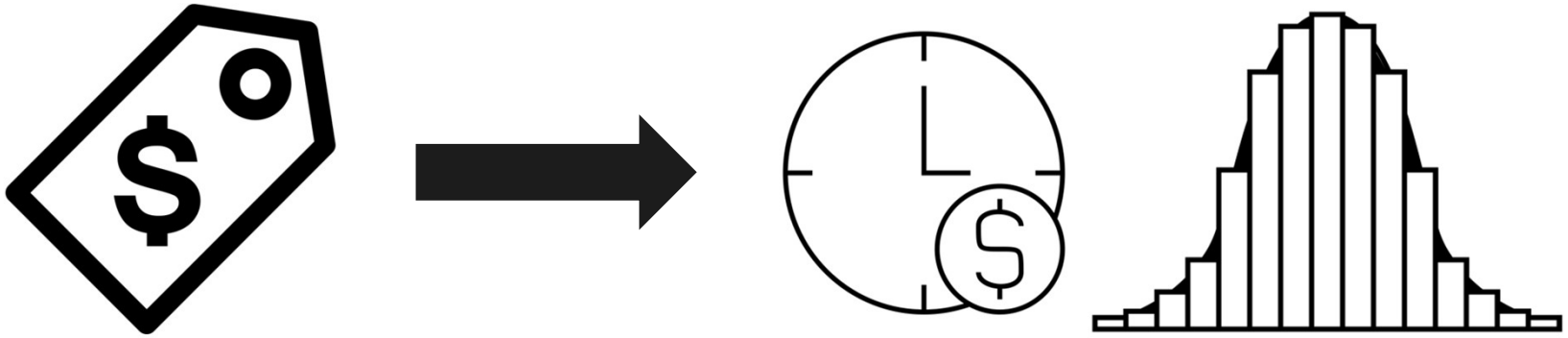
Level of Aggregation: Digital Credit Data

Topic	Provider/Product	Account	Loan	Transaction
Loan contracts	Average loan size, contracted tenure (when fixed)	(Consumer weighted) average loan size, Distribution of number of loans and loan size	Distribution of loan sizes, tenure, contracted APR	Effective tenure
Pricing and fees	Total cost and per loan cost		APR, Distribution of APR	Effective APR
Repayment behavior	Total value defaulted and outstanding loans		Late repayment, default, rollovers	Detailed repayment behavior: early repayment
Multiple borrowing		Multiple account holding	Multiple borrowing	Loan repayment as a function of taking second loan
Statistics:	Just means! Oh no	Means, SDs, Account Level Regressions	Means, SDs, Account Level Regressions	

Pricing: From Provider Level Data to Loan Level

- Data at the product level will include the **total disbursed** and the **total fees charged**
- This means we can find **total expenditures** on credit and **the cost per dollar disbursed**, but not the **APR**

- Data at the loan level will include the **tenure** of loans in addition to the disbursement value and the fees charged
- Can find the **APR**
- Can also find the **distribution** of expenditures, cost per dollar, or APR



Legal, Ethical, and Logistical Considerations

A non-exhaustive list

1. What are the relevant Data Protection Laws in this jurisdiction?
2. Informed consent is not always obtained when using administrative data, is the data collected in line with IPA's IRB?
3. How large/rich/complex do you anticipate the data to be?
4. Who from the partner will be handling the data?

Partnerships

Providers, Regulators, and Researchers

Getting what you ask for

Three main tenants:

1. Make it work for the people delivering the data

1. Build in feedback from the data owner (and/or partner)

1. Trust and security are as important as the analysis

Make it work for the people delivering the data

- If compiling the data is arduous, your partners may not comply fully with the request (same for de-identification, transfer, etc.)
- The amount of effort needed to complete the request may get *lost in translation* when you're not communicating with the IT or data folk

While it's tempting to ask for everything, there is a wisdom to keeping it simple

Build in feedback from the data owner

- If there's a problem with a data request, you want to know
- It's very difficult to validate this data, if the request makes a variable different to compile, this could impact the findings of your research!
- *EX: For our prospective data request with the Central Bank of Nigeria we've built in a window for comment and scheduled information sessions*

Building on trust with security

- Long term investment in relationships with regulators in Kenya
- PII needed to be de-identified by CAK
- Built R scripts that dealt with both direct and indirect identifiers
  - Deleted or hashed direct identifiers (MSISDN)
  - Coarsened indirect identifiers (location, date of birth)
  - This made de-identification by CAK possible, if not perfectly smooth
- Documented security measures taken around transfer and storage of data

```
#### 3.1.1 Input Format ####
## Specify file format as xlsx, xls, csv, or txt
file_format = "csv"

#### 3.1.2 Output Format ####
## Specify output file format as xlsx or csv
output_file_format = "csv"

#### 3.1.3 Filenames ####
## Add the file name(s) in quotes below, separated by commas, without the
file extension. If there is only one dataset, then only include one entry in
the list.
#filenames = c("")

## Also, I have written a function read_filenames() that automatically
collects all files with the format specified above. To use, uncomment the
function below and comment the line above. Please situate the files in a
directory (input_path) with has no other data in it.
filenames = read_filenames(input_path = input_path,file_format = file_format)

#### 3.2 Variables ####

#### 3.2.1 Variable Names ####
## Please add desired variable names in quotes, seperated by commas in the
order they appear in the data. There should be one name for each column in
the data. Please be sure to title variable containing date of birth "dob" and
variable containing phone number "msisdn"
new_names = c("row_number", "account_id", "msisdn", "dob", "gender",
"loan_id", "type", "interest_rate","credit","debit","datetime")

#### 3.2.2 Variables to Drop ####
## These are variables that contain personally identifiable information. This
should include "dob" and "msisdn" if present as well as other sensitive data
(e.g., account number).
dropvars = c("msisdn","dob","account_id")

#### 3.2.3 Date of Birth ####
## Please specify the date format for date of birth by writing a three letter
string where y is year, m in month and, d is day. It should look like one of
the following: ymd, ydm, mdy, myd, dmy, or dym. Note: if the month is written
in text format please replace "m" with "b" for abbreviated months ("Jan"),
and "B" for full month names, e.g., "ymd" becomes "ybd" or "yBd"
date_format = "ymd"
```

Next Step in De-Identification: Hashing App

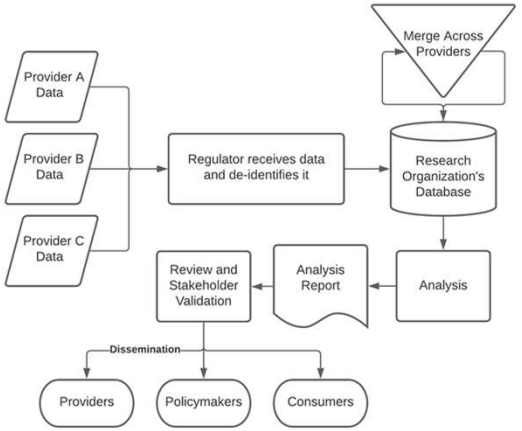
Painless De-Identification

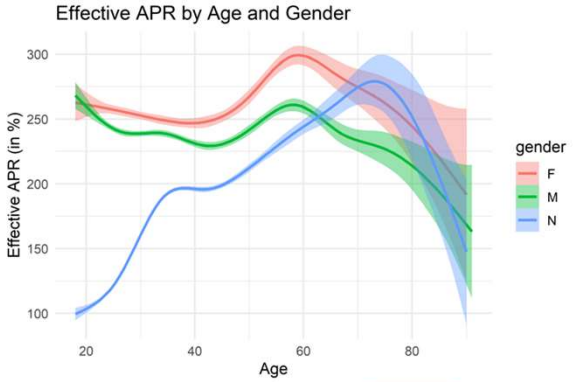
1. Our colleague at CAK learned how to run and edit scripts in R – not easy to do over Skype!
2. Staffing: production level code needs a dedicated programmer
3. What's next: a **hashing app** that we can use for future collaborations

Previewing the Toolkit

Digital credit data analysis: A toolkit for consumer protection supervision

- Tools for the data request:
  - De-identification
  - Data transfer set-up
  - Request letters & templates
  - Sample data (?)
- Analysis & research design:
  - Market monitoring
  - Prediction





national_id	name	gender	dob	msisdn	occupation	latitude	longitude	provider	loan_id	transaction_type	transaction_date	debit	credit
10014322	Andrew Evans		8581	741684528		34.792766	-136.495238	0	08c3d771613e6f7adff	repayment	4/29/2018		1452.7
10025394	Abhishekha Haasenritter		3831	994738809	clerk	-53.22426744	-116.2363021	0	e46b44e777993ddd4c	disbursement	4/5/2018	1004.5	
10047851	Dominic Fonseca		-7256	701652150	spy			0	1461fe41da6d700b2f:	repayment	4/10/2018		1312.71
10047851	Dominic Fonseca		-7256	701652150	spy			0	42a657baf4eac1c116c	disbursement	4/18/2018	885.6769	
10078838	Cabrion Shaw	Male	-1985	152641340	clerk			0	05a4188082784028f8	repayment	4/26/2018		984.016
10108808	Keyva Warrior	Female	1971	775652305	farmer	41.4141525	117.968828	0	653ba06a7e0712ccce	repayment	4/11/2018		1145.8
10190155	Simon Yoo			370788227		-75.66121001	-147.1342263	0	8fad6cebddb4d082e3	disbursement	4/23/2018	1298.573	
10193282	Savannah Estes		-3954	688807323	spy			0	c00a99f05a491ef9689	repayment	4/17/2018		1228.5C
10193472	Sydney Childs	Female	-12591	179763094	farmer			0	0b4d3a474b6068e007	disbursement	4/21/2018	1143.08	
10244130	Jamie Privett	Female	934	495406030	student	29.95984018	83.89880849	0	6ea7d959dc17ca7637	repayment	4/26/2018		1452.99
10259485	Markangelo Burke	Male	6825	900545206	spy			0	60ba1d91ff5bea4eb7	disbursement	4/2/2018	1135.7	
10286678	Saalima el-Asad	Female	-13481	366663060	clerk	68.05813664	101.734866	0	2e34a35e2c24e182bf	disbursement	4/16/2018	785.15	
10295991	Alexis Perez	Female		424125600	soldier	54.5812179	165.282648	0	8951731b4e017d3f9a	disbursement	4/4/2018	1313.807	
10320111	Timothy Merysca	Male		648400937	merchant	43.33190366	-110.1897526	0	1c1fa33b4a5d0dd97d	repayment	4/23/2018		1974.47
10354329	Ailyn Quintana	Female	-4793	707566582	student	11.31169881	128.2048059	0	b3c43cb8666eea4d63	disbursement	4/30/2018	1109.46	
10420825	Keelan Lawton		-431	344869076	clerk	27.59282396	40.50153282	0	7400350aab5b8af34d	repayment	4/19/2018		1738.18
10429062	Armando Bermudez	Male	-6695	899620233	soldier	-52.94587229	-7.171002561	0	a193a8c7c3471d968e	repayment	4/12/2018		1743.22
10457113	Justin Speier	Male	-8505	904314941	soldier			0	37dee3901dfd203db5	disbursement	4/25/2018	924.55	

COMPETITION AUTHORITY OF KENYA

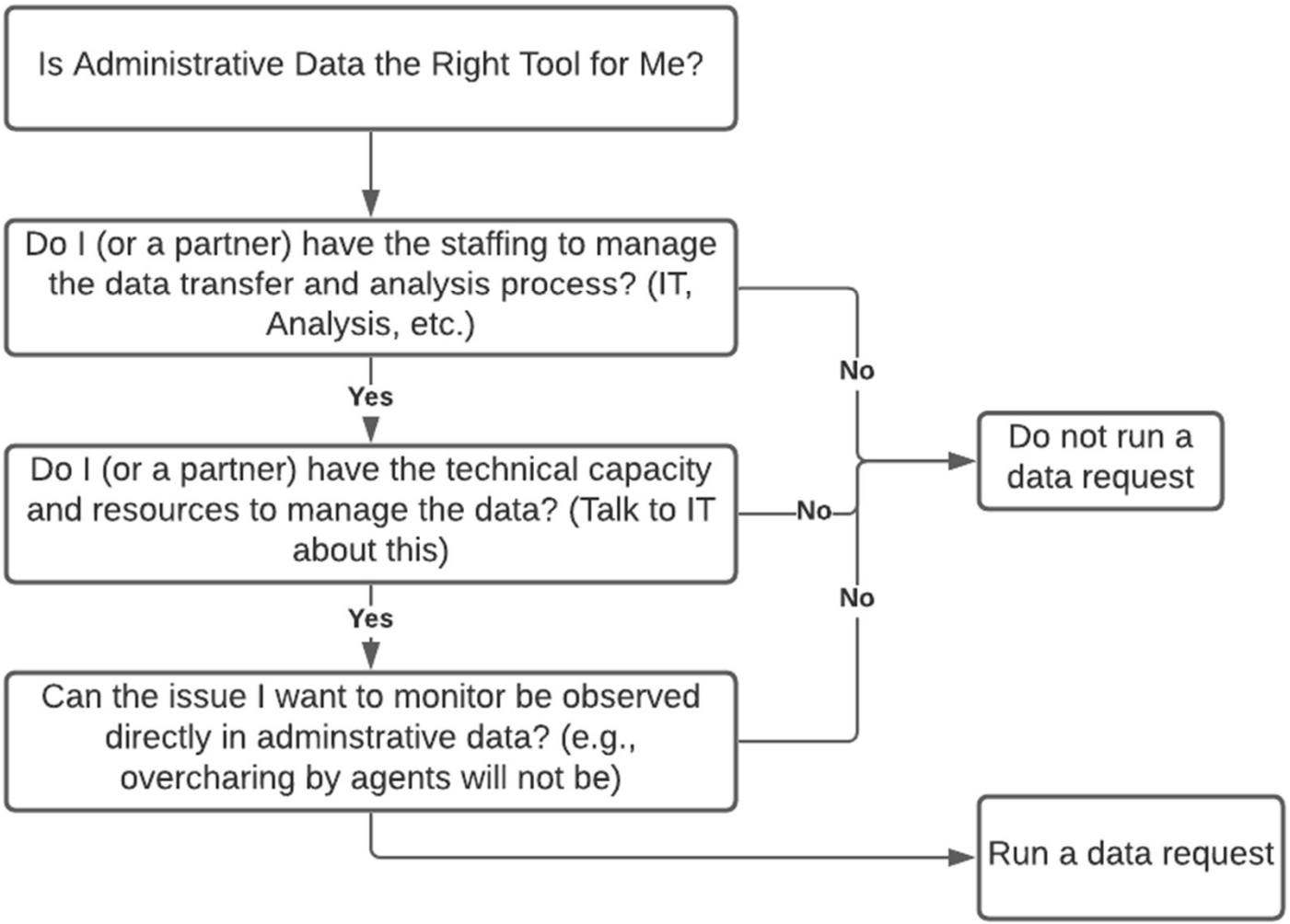
Creating efficient markets for consumers

Discussant: Dr. Adano W. Roba

Director, Planning, Research, Policy & Quality Assurance

Competition Authority of Kenya

2

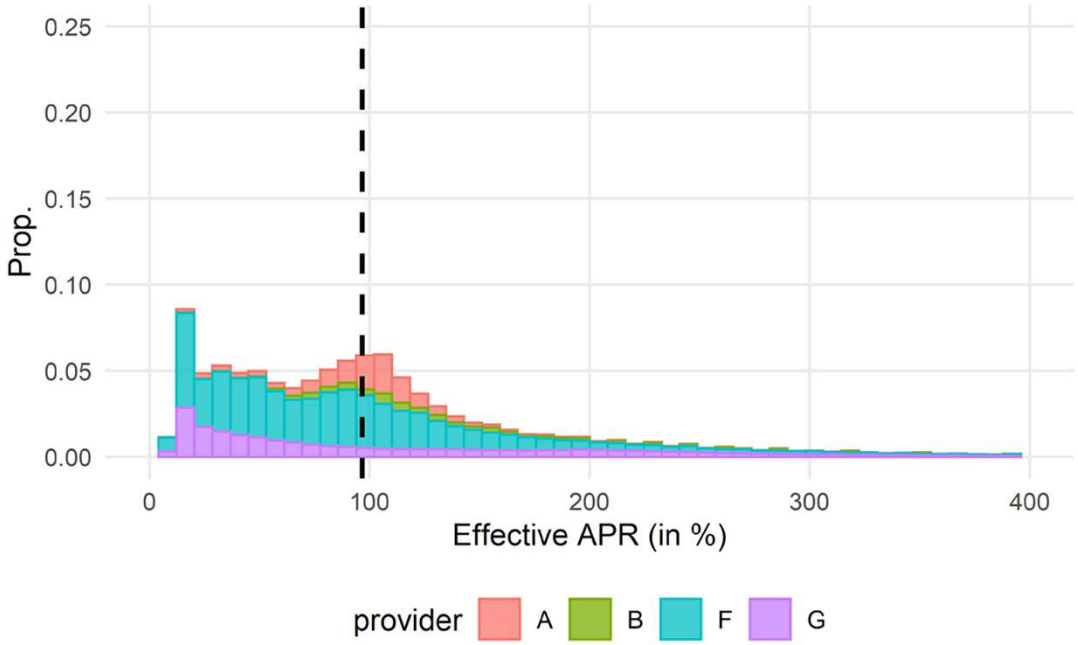

Is Administrative Data the Right Tool for Me?
Do I (or a partner) have the staffing to manage the data transfer and analysis process? (IT, Analysis, etc.)
Yes
Do I (or a partner) have the technical capacity and resources to manage the data? (Talk to IT about this)
Yes
Can the issue I want to monitor be observed directly in adminstrative data? (e.g., overcharing by agents will not be)
No
No
No
Do not run a data request
Run a data request

Slide 35

1 @tjaluka@poverty-action.org here's the flowchart, which I didn't love. Feeling short on time to make it nicer...
Daniel Putman, 10/28/2021

2 Also flowcharts have a bit of a feeling of being "all-inclusive" which I'm a little hesitant to claim
Daniel Putman, 10/28/2021

Using Effective APR to Measure Cost of Credit



- APR captures the annualized cost of all fees related to a loan. We use effective APR, which uses the actual time a borrower paid back the loan in as the measurement of loan tenure.
- Early repayment can make the effective cost of digital credit quite high for some borrowers

Using Effective APR to measure credit prices

Contracted APR	(Actual) APR	Effective APR
$= \left(\frac{Normal\ Fees}{Loan}\right) \times \left(\frac{365\ days}{Tenure}\right) \times 100\%$	$= \left(\frac{Normal + Conditional\ Fees}{Loan}\right) \times \left(\frac{365\ days}{Tenure}\right) \times 100\%$	$= \left(\frac{Normal + Conditional\ Fees}{Loan}\right) \times \left(\frac{365\ days}{Actual\ Days\ to\ Repayment}\right) \times 100\%$

Effective APR responds to early repayment

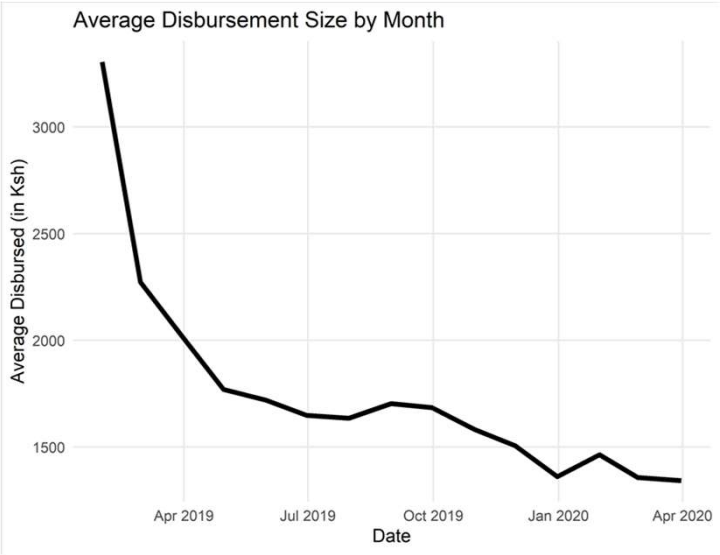
Angie, Benny, Chloe, and Danny each get a 30-day loan with a 10% fee. Angie and Benny repay early, Chloe and Danny repay late.

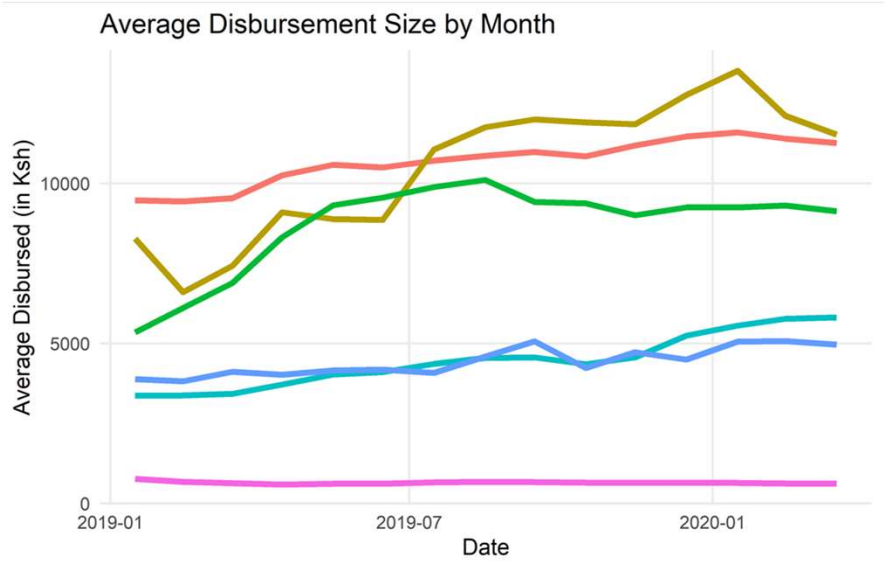
Borrower	Days to repayment	Provider adjustment	Contracted APR	Actual APR	Effective APR
Angie	15 (Early)	None	122%	122%	243%
Benny	30 (On time)	None	122%	122%	122%
Chloe	45 (Late)	Late fee (10%)	122%	243%	162%
Danny	60 (Late)	Rollover (10%)	122%	243%	122%

Digital credit provider administrative data

Volumes and values shifted significantly in 2019

Most providers increased their average loan size, yet average loan size reduced overall—due to large volume of tiny loans disbursed by H2.





Did providers offer larger loans to differentiate themselves or did "small" borrowers substitute into H2 from other sources?

Digital credit provider administrative data

Most multiple account holders also multiple borrow

- Multiple borrowers counted as those who ever borrow from a different provider within 30 days of their previous loan

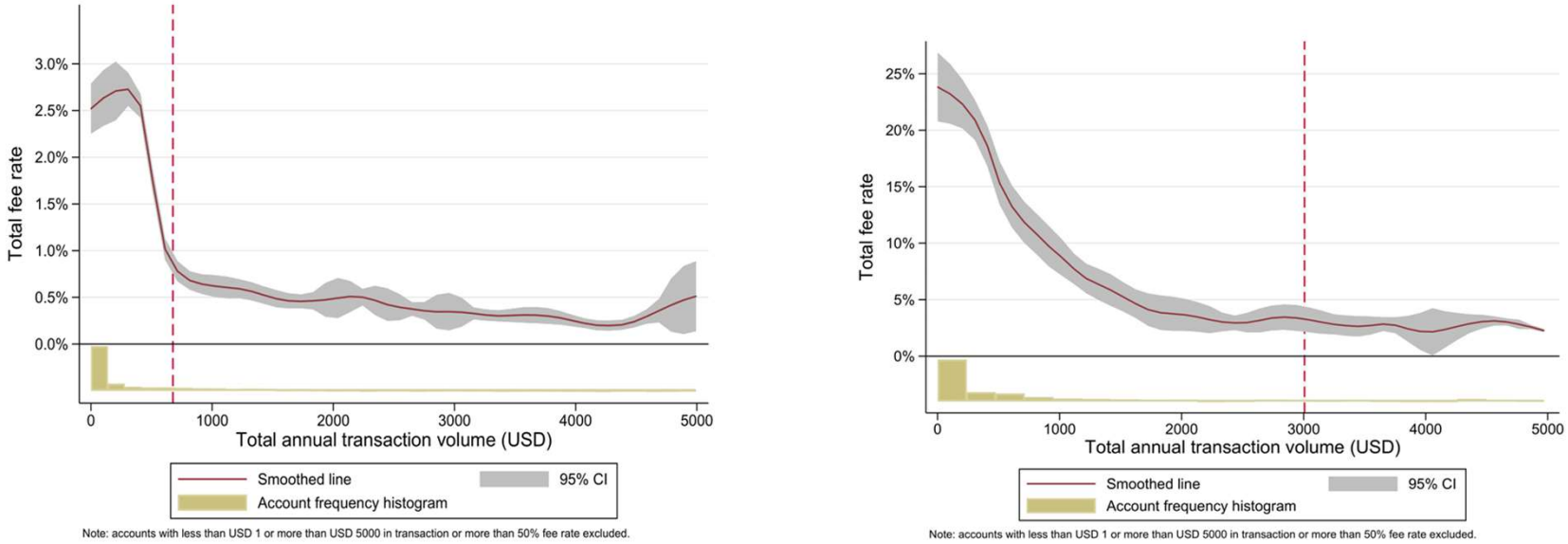
Digital credit provider administrative data

Multiple borrowers default often, early borrowers do not

- Risks related to early borrowing and multiple borrowing may be different
- Multiple borrowing does carry the risk of eventual default
- Early borrowers are good at repaying, but face a high cost of credit

Primary findings from deposit accounts analysis:

2. Costs vary substantially for current and savings accounts



Total Fee Rate for Savings (Left) and Current Accounts (Right)

Current account customers pay fees equaling more than 20% of the volume transacted on their current accounts versus 2.5% for savings accounts, meaning **current account fees are approximately 8 times higher than in savings accounts for similar values.**