Journal of Development Economics 161 (2023) 103016

Contents lists available at ScienceDirect

Journal of Development Economics

journal homepage: www.elsevier.com/locate/devec

Regular article

Program targeting with machine learning and mobile phone data: Evidence from an anti-poverty intervention in Afghanistan ☆

Emily L. Aiken [a], Guadalupe Bedoya [b], Joshua E. Blumenstock [a,*], Aidan Coville [b]

[a] School of Information, University of California, Berkeley, United States of America
[b] Development Impact Evaluation Department, World Bank, United States of America

ARTICLE INFO

JEL classification:
I32
I38
O12
O38
C55

Keywords:
Targeting
Machine learning
Mobile phone data
Afghanistan

ABSTRACT

Can mobile phone data improve program targeting? By combining rich survey data from a "big push" anti-poverty program in Afghanistan with detailed mobile phone logs from program beneficiaries, we study the extent to which machine learning methods can accurately differentiate ultra-poor households eligible for program benefits from ineligible households. We show that machine learning methods leveraging mobile phone data can identify ultra-poor households nearly as accurately as survey-based measures of consumption and wealth; and that combining survey-based measures with mobile phone data produces classifications more accurate than those based on a single data source.

1. Introduction

Each year, hundreds of billions of dollars are spent on targeted social protection programs. The importance of these programs increased dramatically in the wake of the COVID-19 pandemic: In 2020, global extreme poverty increased for the first time in two decades, and most countries expanded their social protection programs, with more than 1.1 billion new recipients receiving government-led social assistance payments (Gentilini et al., 2020).

Determining who should be eligible for program benefits — *targeting* — is a central challenge in the design of these programs (Hanna and Olken, 2018; Lindert et al., 2020). In high-income countries, targeting frequently relies on tax records or other administrative data on income. In low- and middle-income countries (LMICs), where a large fraction of the workforce is informal, programs often require primary data collection. The difficulty and cost of collecting data, and the variable quality of what gets collected, can introduce significant errors in the targeting process (Deaton, 2016; Jerven, 2013; Grosh et al., 2022). These issues are exacerbated in fragile and conflict-affected countries, where two thirds of the world's poor are expected to reside by 2030 (Corral et al., 2020).

This paper evaluates the extent to which non-traditional administrative data, processed with machine learning, can be used for program targeting. Specifically, we match call detail records (CDR) from a large mobile phone operator in Afghanistan to household survey data from the Afghan government's Targeting the Ultra-Poor (TUP) anti-poverty program. Eligibility for the TUP program was determined through a *hybrid targeting method*, combining a community wealth ranking (CWR) and a short follow-up survey. Our analysis assesses the accuracy of three counterfactual targeting approaches at identifying the actual beneficiaries of the TUP program: (i) our *CDR-based method*, which applies machine learning to data from the mobile phone company; (ii) an *asset-based wealth index*, which uses asset ownership to approximate poverty; and (iii) *consumption*, a common benchmark for measuring poverty in LMICs.

Our analysis produces three main results. First, by comparing errors of inclusion and exclusion using the program's hybrid method as a

☆ We thank Seungmin Lee, Maria Camila Ayala and Thomas Escande for excellent research assistance. This work was supported by Defense Advanced Research Projects Agency, United States of America and NIWC under contract N66001-15-C-4066, the NSF, United States of America under grant IIS-1942702, and by the World Bank's Knowledge for Change Program. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright notation thereon. The views, opinions, and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense, the U.S. Government, the World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

* Corresponding author.
E-mail address: jblumenstock@berkeley.edu (J.E. Blumenstock).

https://doi.org/10.1016/j.jdeveco.2022.103016
Received 27 June 2022; Received in revised form 20 November 2022; Accepted 23 November 2022
Available online 28 November 2022

benchmark, we find that the CDR-based method is nearly as accurate as the commonly-employed asset and consumption-based methods for identifying the phone-owning ultra-poor households. Second, we find that methods combining CDR data with measures of assets and consumption are more accurate than methods using any single data source. Third, we find that when non-phone-owning households are included in the analysis, the CDR-based method remains accurate if non-phone-owning households are classified as ultra-poor; however, targeting performance is quite poor if households without phones are ineligible for benefits. After presenting these main results, we compile data from several existing targeting programs to give an indication of the substantial reduction in marginal costs associated with CDR-based targeting.

These results connect two distinct strands of prior work. The first is a literature on program targeting, which studies the effectiveness of different mechanisms for identifying program beneficiaries. In LMICs, research has focused on the performance of proxy means tests (PMTs) (Grosh and Baker, 1995; Filmer and Pritchett, 2001; Brown et al., 2018), community-based targeting strategies (CBTs) (Alatas et al., 2012; Fortin et al., 2018), and related approaches (Banerjee et al., 2007; Karlan and Thuysbaert, 2019; Premand and Schnitzer, 2020). A meta-analysis by Coady et al. (2004), which includes 8 PMTs and 14 community-based programs, finds little difference in targeting accuracy between the two methods — but notes that targeting is regressive in a quarter of programs reviewed. In addition to issues with targeting accuracy, the current methods available for poverty targeting in LMICs are time- and resource-intensive, and may be infeasible in fragile or conflict-affected areas or in contexts where social interaction is limited, such as during a pandemic.

The second body of work explores the extent to which non-traditional sources of data, in conjunction with machine learning, might help address data gaps in LMICs (e.g., Blumenstock, 2016; Burke et al., 2021). Much of this work focuses on estimating the geographic distribution of poverty at fine spatial granularity, using data from satellites (Jean et al., 2016; Engstrom et al., 2017), mobile phones (Blumenstock et al., 2015; Hernandez et al., 2017), social media (Fatehkia et al., 2020; Sheehan et al., 2019), or some combination of these data sources (Steele et al., 2017; Pokhriyal and Jacques, 2017; Chi et al., 2022). Most relevant to our current analysis, two prior papers investigate whether the mobile phone use can approximate the wealth of individual mobile subscribers. Blumenstock et al. (2015) show that CDR data are predictive of an individual-level asset-based wealth index among a nationally representative sample of 856 Rwandan mobile phone owners ($r = 0.68$). Blumenstock (2018b) finds similar results with a sample of 1234 male heads of households in the Kabul and Parwan districts of Afghanistan. While these results show that phone data can be used to *predict* poverty levels, they do not evaluate whether those poverty estimates are of sufficient quality for real-world policy applications.

Our paper connects these two literatures by rigorously assessing the extent to which phone-based estimates of poverty can help with program targeting (Blumenstock, 2020). We believe the analysis will be especially relevant to the increasing number of interventions that rely on mobile money to distribute cash payments (Gentilini et al., 2020), and the growing number of contexts where mobile phone data are being made available for humanitarian purposes (Milusheva et al., 2021). For example, in just the past few years, mobile money was used to make cash transfer payments in countries including Bangladesh (Ali and May, 2021), Ghana (Karlan et al., 2021), Liberia (USAID, 2021), and Malawi (Paul et al., 2021). Mobile phone *data* has been used to guide cash transfers in Colombia (Gentilini et al., 2020), the Democratic Republic of the Congo (Gentilini et al., 2021), Pakistan (Gentilini et al., 2020), and Togo (Aiken et al., 2022).[1]

[1] The anti-poverty program implemented in Togo and described by Aiken et al. (2022) was based on the methods developed and evaluated in this paper. Due to the time-sensitive nature of the COVID-19 response described in Aiken et al. (2022), the two academic articles are in circulation concurrently.

The context of our empirical analysis – identifying ultra-poor households in Afghanistan – is a particularly challenging environment for data collection and program targeting, as 62% of the households classified as not *ultra*-poor still fall below the national poverty line. In such environments, when traditional options for targeting are not feasible, these methods may provide a viable alternative for identifying households with the greatest need. Given the policy relevance of these results, we conclude our analysis by discussing important ethical and logistical considerations that may influence how CDR methods are used to support targeting efforts in practice.

2. Data and methods

2.1. Targeting the 'ultra-Poor'

Our empirical analysis relies on survey data collected as part of the Targeting the Ultra-Poor (TUP) program implemented by the government of Afghanistan with support from the World Bank. The TUP program was a "big push", providing multi-faceted benefits to 7,500 ultra-poor households in six provinces of Afghanistan between 2015 and 2018 (Bedoya et al., 2019). Our analysis uses data from the baseline and targeting surveys from an impact evaluation of the TUP program conducted in Balkh province.

Ultra-Poor designation. Eligibility for the TUP program was determined based on geographic criteria,[2] followed by a two-step process including a community wealth ranking (CWR) and a follow-up in-person survey. CWRs were conducted separately in each village, coordinated by a local NGO and village leaders, in collaboration with the government team. The CWR was followed by an in-person survey to determine whether nominated households met a set of qualifying criteria, coordinated by the NGO and government representatives, and based on a measure of multiple deprivation.

For a household to be designated as *ultra-poor*, and therefore eligible for program benefits, it had to be considered extreme-poor in the CWR (43% of households), and also meet at least three of six criteria:

1. Financially dependent on women's domestic work or begging
2. Owns less than 800 square meters of land or living in a cave
3. Primary woman under 50 years old
4. No adult men income earners
5. School-age children working for pay
6. No productive assets

Ultimately, 11% of the households classified as extreme-poor in the community wealth ranking step — 6% of the total population in the study villages — were classified as ultra-poor and were thus eligible for TUP benefits.

2.2. Household surveys

To facilitate Bedoya et al.'s (2019) impact evaluation of the TUP program, household surveys were conducted in 80 of the poorest villages of Balkh province. A total of 2852 households were surveyed, with ultra-poor households (N=1173) oversampled relative to non-ultra-poor households (N=1679).[3] Surveys were conducted between February and April 2016, following the CWR and eligibility verification. This survey window was timed to occur in the late winter and

[2] The poorest villages were identified by the availability of veterinary services, financial institutions, and social services, and being relatively accessible (Bedoya et al., 2019).

[3] In our analysis, we restrict to the 2814 households for which asset and consumption data are nonmissing.

early spring, a few months before the harvesting season for wheat in Balkh.

The household survey was a long-form in-person survey that took approximately 3 h for each household to complete. The survey covered a wide range of topics, including several modules related to household poverty and deprivation that feature in our analysis.

Consumption. The consumption module of the TUP survey captured information on household food consumption for the week prior to the interview and non-food expenditures for the month or year prior to the interview. These are used to construct monthly *per capita* consumption values, as detailed in Bedoya et al. (2019). Based on these data, we measure the logarithm of *per capita* monthly consumption, using the same approach that the Afghan government used to determine the national poverty line. This monthly consumption aggregate thus captures a short-term (weekly) measure of food consumption during one of the planting seasons, as well as a medium-term (monthly and annual) measure of non-food expenditures (Deaton, 1997; Ravallion, 1998).

Asset index. We use survey data on household assets to construct a wealth index for each household, which provides an indication of each household's wealth relative to others in the survey. Specifically, we calculate the first principal component of variation in household asset ownership based on the sixteen items listed in Table S1, across the 2814 households with complete asset data, after standardizing each asset variable to zero mean and unit variance. This wealth index explains 25.3% of the variation in asset ownership. Figure S1 shows the distribution of the underlying asset index components and Table S1 shows the direction of the first principal component. Broadly, we expect that the asset index will provide an indication of each household's long-term economic status, relative to other households in the survey.

Other variables. The TUP surveys collected several other covariates that we use in subsequent analysis. These include a food security index (composed of variables relating to the skipping and downsizing of meals, separately for adults and children), a financial inclusion index (composed of access to banking and credit, knowledge of banking and credit, and savings), and a psychological well-being index for the primary woman (standardized weighted scores on the Center for Epidemiological Studies Depression scale, the World Values Survey happiness and satisfaction questions, and Cohen's Stress Scale) – see Bedoya et al. (2019). The survey also collected data from each household on mobile phone ownership. Nearly all (99%) households with a cell phone provided their phone numbers and consented to the use of their call detail records for this study.

Sample representativity. Portions of our analysis are restricted to the 535 households from the TUP survey with phone numbers that match to our CDR (see Section 2.3). Table 1 and S2 compare characteristics of these households to the full survey population. There are some systematic differences: the 535-household sample is wealthier, which is consistent with households in the subsample being required to own at least one phone. For instance, while 88% of non-ultra-poor households in the TUP survey own at least one phone, only 72% of ultra-poor households own at least one phone.

Comparing survey-based measures of well-being and deprivation. As shown in Table 1 and Figure S3, the two survey-based measures of well-being are only weakly correlated. In the full sample, the correlation between the asset index and consumption is just 0.37; in the matched subsample, the correlation is 0.34. These modest correlations may be due in part to the fact that, as discussed above, the consumption data capture short- and medium-term deprivation, whereas the asset index is a better indicator of long-term wealth. Measurement error may also weaken these empirical correlations.

Also notable is the weak relationship between the two survey-based measures of deprivation and the ground truth ultra-poor designation: while the ultra-poor population makes up 27% of the overall subsample, less than half of the ultra-poor fall into the bottom 27% of the sample by wealth index or consumption. These differences may be partly attributable to measurement error, but they surely also arise from the fact that they are conceptually distinct constructs: while the consumption and asset indices focus primarily on economic flows and stocks, respectively, the ultra-poor designation was designed to be more holistic and multidimensional, informed in part by community perceptions of vulnerability (Sen, 1992; Alkire et al., 2015).

The fact that the ultra-poor designation is not strongly correlated with the survey measures of consumption and wealth has important implications for the targeting analysis presented below. In particular, it suggests – and our later results affirm – that a policy targeted solely on assets or consumption data will do a poor job of differentiating between ultra-poor and non-ultra-poor. The relatively weak correlation between consumption and the asset index also hints at a later finding that targeting based on a combination of the two data sources performs better than targeting on a single source in isolation.

Sample weights. Since the TUP survey oversampled the ultra-poor (by a factor of roughly 12), portions of our analysis use sample weights to adjust for population representativeness. When sample weights are applied, it is explicitly noted; if not mentioned, no weights are applied. After sample weights are applied, the ultra-poor make up 5.98% of the overall population, and 4.63% of our matched subsample.

2.3. Mobile phone metadata

In a follow-up survey conducted in 2018, we requested informed consent from survey respondents to obtain their mobile phone CDR and match them to the survey data collected through the TUP project. CDR contain detailed information on:

- **Calls:** Phone numbers for the caller and receiver, time and duration of the call, and cell tower through which the call was placed
- **Text messages:** Phone numbers for the caller and recipient, time of the message
- **Recharges:** Time and amount of the recharge

For participants who consented, we match baseline survey data (collected November 2015–April 2016) to CDR covering that same period, obtained from one of Afghanistan's main mobile phone operators. For households with multiple phones and a designated household head (N=65), we match to CDR for the phone belonging to the household head. For households where the household head does not have a phone and someone else does (N=17), we match to CDR for one of the households' phones selected at random. In total, for the 535 households in our sample, 629,543 transactions took place in the months of November 2015 to April 2016, broken down into 310,883 calls, 305,756 text messages, and 12,904 recharges.

From these CDR, we compute a set of 797 behavioral indicators that capture aggregate aspects of each individual's mobile phone use (de Montjoye et al., 2016). This set includes indicators relating to an individual's communications (for example, average call duration and percent initiated conversations), their network of contacts (for example, the entropy of their contacts and the balance of interactions per contact), their spatial patterns based on cell tower locations (for example, the number of unique antennas visited and the radius of gyration), and their recharge patterns (including the average amount recharged and the time between recharges). The distributions of a sample of these indicators are shown in Figure S4.

Table 1
Summary statistics for different samples of respondents

Outcome	(1) Full sample (All observations)	(2) Matched Subsample	(3) Unmatched Owns Phone	(4) Unmatched No Phone
Panel A: Balance of Covariates				
Ultra-Poor	0.42 (0.49)	0.27 (0.45)	0.40 (0.49)	0.66 (0.47)
Asset Index	0.00 (2.01)	1.36 (2.60)	−0.05 (1.76)	−1.35 (0.79)
Log Expenditures	4.43 (0.71)	4.64 (0.70)	4.46 (0.70)	4.12 (0.65)
# Phones	1.35 (1.18)	1.72 (1.33)	1.59 (1.04)	0.00 (0.00)
Food Security Index	0.30 (0.90)	0.35 (0.74)	0.34 (0.93)	0.10 (0.89)
Financial Inclusion Index	0.15 (1.27)	0.34 (1.39)	0.15 (1.32)	−0.05 (0.79)
Psychological Well-being Index	0.35 (1.01)	0.38 (1.00)	0.43 (0.97)	−0.02 (1.07)
CWR Group	0.62 (0.90)	0.89 (1.02)	0.62 (0.88)	0.26 (0.66)
Panel B: Correlations Between Outcomes				
Ultra-Poor ⟷ Asset Index	−0.32	−0.30	−0.27	−0.14
Ultra-Poor ⟷ Consumption	−0.39	−0.30	−0.39	−0.26
Asset Index ⟷ Consumption	0.37	0.34	0.34	0.15
N	2814	535	1807	472

Notes: Table reports average characteristics, with standard deviations in parentheses, of TUP survey respondents. Each column represents a different sample of respondents: (1) all respondents in the TUP survey; (2) Just those respondents who own a phone, where the phone number matches to the CDR obtained from the mobile phone operator; (3) Respondents who report owning a phone, but whose phone number does not match to the CDR obtained from the operator; (4) Respondents who report they do not own a phone.

2.4. Machine learning predictions

CDR-based method. Extending the approach described in Blumenstock et al. (2015), we test the extent to which ultra-poor status can be predicted from CDR. This analysis uses the 535 TUP households who match to CDR to train a supervised machine learning algorithm to predict ultra-poverty status from the mobile phone features. The intuition — also highlighted in Figure S4 — is that ultra-poor individuals use their phones very differently than non-ultra-poor individuals, and machine learning algorithms can use those differences to predict ultra-poor status.

Our main analysis uses a gradient boosting model, which generally out-performs several other common machine learning algorithms for this task (see Table S3). The feature importances for the trained model are shown in Table S2. To limit the potential for overfitting, probabilistic predictions are generated via 10-fold cross-validation, with folds stratified to preserve class balance.[4] Additional details on the machine learning methods are provided in Appendix A.

Combined methods. We also evaluate several approaches that use data from multiple sources to predict ultra-poor status. Our main *combined method* trains a logistic regression to classify the ultra-poor and non-ultra-poor households using the predicted ultra-poor probability from the CDR-based method (i.e., the output of the gradient boosting algorithm described above), as well as asset and consumption data collected in the TUP survey. For comparison, we similarly evaluate the performance of methods that combine only two of the available data sources (i.e., assets plus consumption, assets plus CDR, and consumption plus CDR). Predictions for each of the combined methods are pooled over 10-fold cross-validation.

2.5. Targeting accuracy evaluation

Evaluation on matched subsample. Our main analysis focuses on the 535 households for which we observe both CDR and survey data, and evaluates whether machine learning methods leveraging CDR data can accurately identify households designated as ultra-poor by the TUP program (using the two-step hybrid approach described in Section 2.1). We compare the performance of the CDR-based method to the performance of methods based on the wealth index, consumption data, and combinations of these data sources.[5] Each targeting method is evaluated based on classification accuracy, errors of exclusion (ultra-poor households misclassified as non-ultra-poor) and errors of inclusion (non-ultra-poor households misclassified as ultra-poor). We focus on the ultra-poor designation as the 'ground truth' status of the household, against which other methods are evaluated, since it is the most carefully vetted measure of well-being for this population, and the proxy that the government used to target TUP benefits.

To evaluate the performance of the CDR-based and combined methods, we pool out-of-sample predictions across the ten cross-validation folds, so that every household in our dataset is associated with a CDR-based predicted probability of ultra-poor status that is produced out-of-sample.[6] To account for class imbalance, we evaluate model accuracy using a "quota method", by selecting a cut-off threshold for ultra-poor qualification such that each method identifies the proportion of ultra-poor households in our subsample; this cut-off also balances inclusion and exclusion errors. This quota-based approach reflects a scenario in which a program has a fixed budget constraint; it is also frequently used in the targeting literature (Alatas et al., 2012; Schnitzer and Stoeffler, 2021). In our 535-household matched dataset this threshold is 27%; in other samples (see following subsection), the percentage is different. We evaluate each method for precision (positive predictive value) and recall (sensitivity). To capture the trade-off between inclusion and exclusion errors for varying values of this threshold, we also construct receiver operating characteristic (ROC) and precision–recall curves for each method and consider the area under the ROC curve (AUC) as a measure of targeting quality. For each evaluation metric (precision, recall, and AUC), we bootstrap 1000 samples from

[4] While cross validation is a standard evaluation strategy in the machine learning literature, for robustness we present results using a basic single train-test split in Table S6.

[5] The CDR-based method uses supervised learning to model the ultra-poverty outcome, whereas the asset- and consumption-based approaches do not. To assess the importance of this difference, we experiment with applying machine learning methods to the asset and consumption data to model the ultra-poverty outcome. In results shown in Table S4, we find that a machine-learned asset predictor provides slight improvements on the standard asset-based wealth index and consumption measures. We continue to use the standard asset and consumption measures as benchmarks in the remainder of the paper, however, as they are the targeting methods most frequently used in practice.

[6] In Table S6, we show that results are unchanged when we use a single train-test split, instead of 10-fold cross-validation.

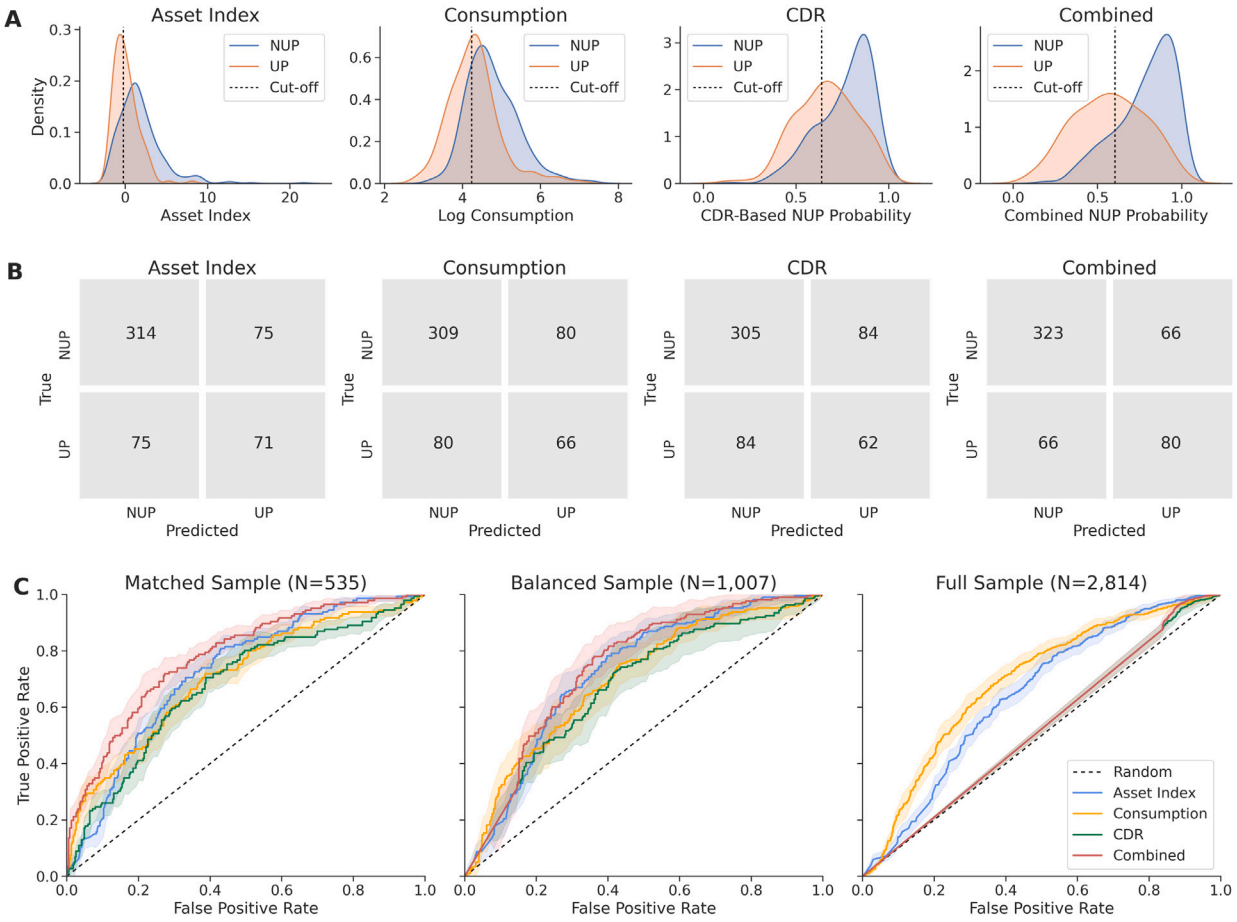

Fig. 1. Predicting ultra-poor status from mobile phone Call Detail Records.
Notes: Panel A: Comparing the predictive accuracy of assets, consumption, and CDR-based methods for identifying the ultra-poor in our 535-household matched sample. To adjust for class balance, thresholds for classification (shown in dashed black vertical lines) are selected such that the correct number of households are identified as ultra-poor. Panel B: Confusion matrices showing the targeting accuracy of each method shown in Panel A. Panel C: ROC curves for each of the four targeting methods. In the third subplot, the CDR-based and combined methods target non-phone-owning households first as described in Section 2.5.

the original dataset to calculate the standard deviation of the mean of the accuracy metric. Each bootstrapped sample is of the same size as the original dataset, drawn with replacement.

Accounting for households without phones. In order to focus our attention on how differences in the data used for program targeting affect targeting performance, our main results are based on the sample of 535 households for whom we have both survey data and mobile phone data. We also present results that show how performance is affected when the analysis includes TUP households for whom we do not have mobile phone data (typically because they do not have a phone or because they use a different phone network than the one who provided CDR). We provide analysis that targets such households (1) before households with CDR, or (2) after households with CDR (see Section 3.4). These results are evaluated on three different samples:

1. *Matched Sample:* The 535 households for whom could match survey responses to CDR.
2. *Balanced Sample:* This sample includes the 535 matched households as well as the 472 households in the TUP survey who report not owning any phone. It excludes households that own a phone on a different phone network than the one who provided CDR. The motivation for this sample is to provide an indication of targeting performance in a regime in which CDR can be used to target all phone-owning households. In addition to applying sample weights from the survey, households that do not own a phone are downweighted so that the balance of phone owners to non-phone-owners (with sample weights applied) is the same as in the baseline survey as a whole (with sample weights applied, 84% phone owners).
3. *Full Sample:* All 2814 households in the TUP baseline survey for which asset and consumption data are available, with sample weights applied.

Note that the quota used to evaluate targeting changes for each sample, based on the number of households that are ultra-poor in the sample. For the matched sample, the targeting quota is 27.29%; for the balanced sample and full sample the quotas are 5.47% and 6.02%, respectively.

3. Results

3.1. Performance of targeting methods

Our first set of results evaluate the extent to which different targeting methods can correctly identify ultra-poor households. This analysis compares the performance of CDR-based targeting methods to asset-based and consumption-based targeting, using the sample of 535 households for which survey data and CDR data are both available.

An overview of these results is provided in Fig. 1. The top panel (Fig. 1a) shows the distribution of assets and consumption, as well as the distribution of predicted probabilities of being non-ultra-poor generated by the CDR-based and combined methods, separately for the ultra-poor and non-ultra-poor. The dashed vertical line indicates the

Table 2
Targeting simulation results.

Targeting method	(1) AUC	(2) Accuracy	(3) Precision	(4) Recall
Panel A: Matched Sample (N=535) - for whom we have survey and CDR data				
Random	0.50 (0.028)	0.60 (0.025)	0.27 (0.038)	0.27 (0.038)
Asset Index	0.73 (0.024)	0.72 (0.020)	0.49 (0.041)	0.49 (0.041)
Consumption	0.71 (0.026)	0.69 (0.023)	0.45 (0.038)	0.45 (0.038)
CDR	0.68 (0.027)	0.69 (0.021)	0.42 (0.042)	0.42 (0.042)
Combined	0.78 (0.022)	0.75 (0.020)	0.55 (0.039)	0.55 (0.039)
Panel B: Balanced Sample (N=1007) - as above, plus households without phones				
Random	0.50 (0.017)	0.90 (0.006)	0.05 (0.010)	0.05 (0.010)
Asset Index	0.72 (0.026)	0.90 (0.006)	0.10 (0.013)	0.10 (0.013)
Consumption	0.70 (0.028)	0.90 (0.006)	0.15 (0.025)	0.15 (0.025)
CDR (Target Phoneless First)	0.68 (0.030)	0.90 (0.006)	0.11 (0.035)	0.11 (0.035)
CDR (Target Phoneless Last)	0.51 (0.028)	0.90 (0.006)	0.12 (0.033)	0.12 (0.033)
Combined (Target Phoneless First)	0.74 (0.026)	0.90 (0.006)	0.11 (0.046)	0.11 (0.046)
Combined (Target Phoneless Last)	0.57 (0.022)	0.90 (0.006)	0.18 (0.007)	0.18 (0.007)
Panel C: Full Sample (N=2814) - as above, plus households with phones on other networks				
Random	0.50 (0.009)	0.89 (0.005)	0.06 (0.007)	0.06 (0.007)
Asset Index	0.65 (0.017)	0.89 (0.005)	0.07 (0.014)	0.07 (0.014)
Consumption	0.69 (0.015)	0.89 (0.006)	0.08 (0.031)	0.08 (0.031)
CDR (Target Phoneless First)	0.52 (0.008)	0.89 (0.005)	0.06 (0.008)	0.06 (0.008)
CDR (Target Phoneless Last)	0.48 (0.008)	0.89 (0.005)	0.08 (0.010)	0.08 (0.010)
Combined (Target Phoneless First)	0.52 (0.008)	0.89 (0.005)	0.06 (0.008)	0.06 (0.008)
Combined (Target Phoneless Last)	0.49 (0.008)	0.89 (0.005)	0.09 (0.009)	0.09 (0.009)

Notes: Four different measures of performance (columns) reported for different targeting methods (rows), using different samples of survey respondents (panels). Standard deviations, calculated using 1000 bootstrap samples, in parentheses. Panel A: The 535-household subsample that is matched to CDR. Panel B: The 535-household matched sample, plus the 472 households that do not have a phone; this is meant to approximate targeting performance if CDR from all mobile networks were available. Sample weights are applied as described in Section 2.5. Panel C: All 2814 observations from the TUP survey, including households matched to CDR, households that own phones not matched to CDR, and households without phones, with sample weights applied. For Panels B and C, we simulate two types of CDR-based targeting: targeting households without phones first and targeting households without phones last.

threshold at which point 27% of households are classified as ultra-poor; we use this quota because 27% of households in this sample were designed as ultra-poor by TUP. Fig. 1b provides confusion matrices that compare the true status (rows) against the classification made by each method (columns). These confusion matrices are also used to calculate the measures of precision and recall reported in Table 2 Panel A.

We find that the CDR-based method (precision and recall of 42%) is close in accuracy to methods relying on assets (precision and recall of 49%) or consumption (precision and recall of 45%). To evaluate the trade-off between inclusion errors and exclusion errors resulting from selecting alternative cut-off thresholds, Fig. 1c shows the ROC curve associated with each classification method. The Area Under the Curve (AUC) scores for these curves, listed in Table 2, are comparable among methods, with assets (AUC=0.73) slightly superior to consumption (AUC=0.71) and the CDR-based method (AUC=0.68). The corresponding Precision–Recall curves are shown in Figure S5.

3.2. Comparison of errors across methods

To better understand the nature of the mis-classification errors arising from the different datasets used for targeting, Table 3 compares the characteristics of correctly and incorrectly classified households for three different methods (targeting on assets, consumption, and CDR). Panel A highlights differences between ultra-poor households correctly classified as ultra-poor (True Positives) and ultra-poor households *mis*-classified as non-ultra-poor (False Negatives, also referred to as exclusion errors). Likewise, Panel B highlights differences between non-ultra-poor households correctly classified as non-ultra-poor (True Negatives), and non-ultra-poor households mis-classified as ultra-poor (False Positives, or inclusion errors). This analysis uses the *matched* sample (see Table 2) to highlight differences that arise when switching from one targeting dataset to another, on a population of households that are observed in all three datasets.[7]

Across methods, false negatives (exclusion errors) have higher levels of food security, financial inclusion, and psychological well-being than true positives – that is, all three targeting methods misclassify ultra-poor households as non-ultra-poor when those ultra-poor households are better-off, according to other observable characteristics not used in the targeting. Likewise, false positives (inclusion errors) tend to fare worse than true negatives across these same measures. The exact pattern of differences depends on the targeting method; for instance, asset-based targeting (first set of columns) tends to misclassify ultra-poor as non-ultra poor when they have assets (the difference of −2.21 is large), but errors are not systematically correlated with consumption (the difference of −0.19 is relatively small). The CDR-based method in particular tends to prioritize households that score low on these alternative measures of well-being. These patterns suggest that the CDR-based targeting method may capture aspects of well-being that are not captured by standard survey-based measures of poverty such as wealth and consumption.

To test for systematic misclassification of certain types of households, Table 4 displays the overlap in errors of exclusion and inclusion between methods. Our results suggest that the three classifiers misidentify the same households at a rate only slightly above random.[8]

[7] Similar analysis could also be performed using the balanced sample or the full sample; however, results would conflate differences caused by the targeting data (the current focus of Table 3) with the differences that arise from considering (or excluding) households without mobile phones (the current focus of Table 2).

[8] The rates of overlap should be interpreted relative to the expected overlap in errors for random classifiers. Based on our selection of thresholds such that 27% of the sample is identified as ultra-poor, our three classifiers misidentify 15%–27% of the non-ultra-poor and 51%–65% of the ultra-poor. If these classifiers were random, we would expect approximately 20% overlap in inclusion errors and 55% overlap in exclusion errors.

Table 3
What types of households are misclassified?

Panel A: Ultra-Poor Households (Differences Between True Positives and False Negatives)									
	Asset Index			Consumption			CDR		
	TP	FN	Diff.	TP	FN	Diff.	TP	FN	Diff.
Ultra-Poor	1.00 (0.00)	1.00 (0.00)	0.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.00 (0.00)
Asset Index	−1.03 (0.49)	1.18 (1.34)	−2.21 (0.17)	−0.34 (1.09)	0.47 (1.69)	−0.81 (0.23)	−0.09 (1.16)	0.25 (1.70)	−0.34 (0.24)
Consumption	4.21 (0.70)	4.40 (0.62)	−0.19 (0.11)	3.78 (0.32)	4.74 (0.56)	−0.96 (0.07)	4.29 (0.60)	4.32 (0.71)	−0.02 (0.11)
# Phones	0.89 (0.68)	1.63 (1.12)	−0.74 (0.15)	1.02 (0.73)	1.48 (1.14)	−0.46 (0.16)	1.18 (0.61)	1.33 (1.21)	−0.16 (0.15)
Food Security Index	−0.59 (1.13)	−0.51 (1.10)	−0.08 (0.18)	−0.83 (1.19)	−0.32 (0.99)	−0.51 (0.18)	−0.51 (1.14)	−0.58 (1.09)	0.07 (0.19)
Financial Inclusion Index	−0.00 (0.79)	0.29 (1.02)	−0.29 (0.15)	0.10 (0.80)	0.19 (1.02)	−0.09 (0.15)	0.16 (0.98)	0.14 (0.88)	0.02 (0.16)
Psychological Wellbeing Index	−0.35 (0.92)	−0.13 (0.94)	−0.22 (0.15)	−0.37 (0.86)	−0.12 (0.98)	−0.24 (0.15)	−0.31 (0.81)	−0.17 (1.02)	−0.14 (0.15)
CWR Group	0.09 (0.44)	0.01 (0.12)	0.07 (0.05)	0.02 (0.12)	0.08 (0.41)	−0.06 (0.05)	0.06 (0.40)	0.04 (0.24)	0.03 (0.06)
Panel B: Non-Ultra-Poor Households (Differences Between True Negatives and False Positives)									
	Asset Index			Consumption			CDR		
	TN	FP	Diff.	TN	FP	Diff.	TN	FP	Diff.
Ultra-Poor	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Asset Index	2.53 (2.62)	−1.08 (0.50)	3.61 (0.16)	2.06 (2.92)	0.94 (1.75)	1.12 (0.26)	1.94 (2.87)	1.43 (2.27)	0.51 (0.30)
Consumption	4.82 (0.66)	4.57 (0.65)	0.25 (0.08)	4.97 (0.58)	3.98 (0.23)	0.99 (0.04)	4.78 (0.68)	4.74 (0.61)	0.04 (0.08)
# Phones	2.11 (1.43)	0.96 (0.76)	1.15 (0.12)	1.98 (1.49)	1.52 (0.92)	0.46 (0.13)	1.91 (1.44)	1.80 (1.24)	0.11 (0.16)
Food Security Index	0.24 (0.87)	−0.16 (1.03)	0.40 (0.13)	0.24 (0.88)	−0.14 (0.99)	0.37 (0.12)	0.15 (0.91)	0.18 (0.94)	−0.02 (0.12)
Financial Inclusion Index	0.80 (4.92)	−0.01 (0.82)	0.82 (0.29)	0.77 (4.94)	0.18 (1.24)	0.59 (0.31)	0.78 (4.98)	0.17 (1.10)	0.61 (0.31)
Psychological Wellbeing Index	0.69 (0.97)	0.21 (0.75)	0.47 (0.10)	0.62 (0.98)	0.49 (0.80)	0.13 (0.11)	0.62 (0.95)	0.49 (0.93)	0.13 (0.12)
CWR Group	1.30 (1.00)	0.84 (0.96)	0.46 (0.12)	1.23 (1.03)	1.13 (0.94)	0.10 (0.12)	1.26 (1.01)	1.01 (0.98)	0.25 (0.12)

Notes: Table shows the average characteristics, with standard deviations in parentheses, of households that are correctly and incorrectly classified by three different targeting approaches (approaches are indicated by column-group headers: Asset Index; Consumption; and CDR), using the matched sample. Panel A highlights differences between ultra-poor households correctly classified as ultra-poor (True Positives, TP) and ultra-poor households mis-classified as non-ultra-poor (False Negatives, FN; i.e., exclusion errors). Panel B highlights differences between non-ultra-poor households correctly classified as non-ultra-poor (True Negatives, TN), and non-ultra-poor households misclassified as ultra-poor (False Positives, FP; i.e., inclusion errors).

Table 4
Overlap in targeting errors between methods.

	Asset Index	Consumption	CDR	Combined
Panel A: Overlap in Errors of Exclusion				
Asset Index	100.00%	65.33%	57.33%	66.67%
Consumption	61.25%	100.00%	56.25%	62.50%
CDR	51.19%	53.57%	100.00%	63.10%
Combined	75.76%	75.76%	80.30%	100.00%
Panel B: Overlap in Errors of Inclusion				
Asset Index	100.00%	26.67%	22.67%	48.00%
Consumption	25.00%	100.00%	16.25%	37.50%
CDR	20.24%	15.48%	100.00%	46.43%
Combined	54.55%	45.45%	59.09%	100.00%

Notes: Table measures the extent to which the targeting errors produced by each pair of targeting methods overlap in the matched sample. Evaluation is performed on the matched sample of 535 TUP respondents. Panel A: Overlap between ultra-poor households that are misclassified as non-ultra-poor (errors of exclusion) for each targeting method. Panel B: Overlap between non-ultra-poor households that are misclassified as ultra-poor (errors of inclusion).

3.3. Combining targeting methods

Since the different targeting methods identify different populations as ultra-poor, there may be complementarities between asset, consumption, and CDR data. As shown in Panel A of Table 2, we find that a *combined method*, which takes as input the wealth index, total consumption, and the output of the CDR-based method, performs better (AUC = 0.78) than methods using any one data source (AUC = 0.68–0.73). As shown in Table S5, the full method also outperforms methods based on any two data sources (AUC = 0.75–0.76). The method that combines CDR and asset data (AUC = 0.76) may, however, be more practical than the combined method, since consumption data is difficult to collect for large populations.

3.4. Targeting households without phones

An important limitation of CDR-based targeting is that households without phones do not generate CDR. Here, we show how targeting performance is impacted when households without phones are included in the analysis. This analysis uses two additional samples of TUP households to evaluate targeting performance: (i) the *balanced sample*, which adds all of the 472 households without phones to the sample of 535 for whom we have matched CDR; the balanced sample is intended to illustrate the performance of CDR-based targeting if CDR were available from all operators in Afghanistan — though it relies on the assumption that phone-owners observed on our mobile network are representative of all phone owners in Afghanistan (an assumption that is not fully satisfied, as shown in Table 1); and (ii) the *full sample*, which includes all 2814 households surveyed in the TUP baseline with complete asset and consumption data; this sample includes an additional 1807 households who report owning a phone, but whose number does not match to any number in the CDR provided to us by the single mobile operator.[9]

Results in Panels B and C of Table 2 show the performance of each targeting approach on the balanced and full sample, respectively.

[9] These 1807 households include households that report owning a phone on a different network (this network is estimated to have around 30% market share in Afghanistan), as well as phones on our network that were not active during the six-month period of CDR that we analyze.

Note that as described in Section 2.5, different targeting quotas are applied for each panel based on the proportion of each sample that is ultra-poor. In the CDR-based and combined approaches, we report performance when the households without CDR are targeted first (i.e. households without CDR are targeted in a random order and then the households predicted to be poorest are targeted until the quota is reached) as well as when households without CDR are targeted last (i.e., after the 535 households with phones are targeted, households without phones are included in a random order until the quota is reached).

Unsurprisingly, these results suggest that CDR-based targeting is not effective when a large portion of the target population does not own a phone (e.g. Panel C of Table 2, where only 16% of the sample has matching CDR). However, when we simulate more realistic levels of phone ownership in Panel B (84% of the households, based on our survey data), CDR-based targeting is once again comparable to asset- or expenditure-based targeting, particularly when households without phones are targeted first (AUC = 0.72, 0.70, 0.68 for assets, consumption, and CDR, respectively). On the other hand, if households without phones are targeted last (for example, if program administrators base targeting wholly on CDR and provide no benefits to any household without a phone), the CDR-based method only improves marginally on random targeting.[10]

3.5. Additional tests and simulations

Our main analysis considers the household head to be the unit of analysis. As described in Section 2.3, this analysis is based on matching survey-based indices to phone data from the household head, which is consistent with the design of the TUP program and the TUP survey sample frame. An alternative approach matches survey data reported by the household head to all phone numbers associated with the household. As shown in Table S7, the predictive accuracy of these models is slightly attenuated relative to the benchmark results (Table S3).

We also explore the extent to which CDR can be used to predict other measures of socioeconomic status. Our main analysis focuses on the household's ultra-poor designation as the ground truth measure of poverty, since this label was both carefully curated and the actual criterion used to determine TUP eligibility. In Table S8, we report the accuracy with which CDR (obtained from the household head, who is typically male) can predict consumption and asset-based wealth (elicited from the primary woman of each household).[11] In general, these machine learning models trained to directly predict consumption or asset-based wealth do not perform well. This result contrasts with prior work documenting the predictive ability of CDR for measuring asset-based wealth (e.g. Blumenstock et al., 2015). We suspect a key difference in our setting – aside from the fact that we are matching CDR to socioeconomic status at the *household* rather than the *individual* level – is the homogeneity of the beneficiary population: whereas Blumenstock et al. (2015) uses machine learning to predict the wealth of a nationally-representative sample of Rwandan phone owners, our sample consists of 535 individuals from the poorest villages of a single province in Afghanistan, where even the relatively wealthy households are quite poor.

4. Discussion

Our key finding is that, in a sample of 535 phone-owning households in poor villages in Afghanistan, machine learning methods leveraging phone data are nearly as accurate at identifying ultra-poor households as standard asset- and consumption-based methods. Further, we find that methods combining survey data with CDR perform better than methods using a single data source. However, as we demonstrate empirically, low rates of phone ownership — or the inability to access data from all operators — can undermine the value of CDR-based targeting. In our setting, the CDR-based approach still works well if households without phones are targeted before the CDR-based algorithm selects the poorest households with phones. However, this approach may not be appropriate in other contexts where phone ownership is less predictive of wealth, or where potential beneficiaries have the ability to strategically under-report phone ownership (Björkegren et al., 2020).

As mobile phone penetration rates continue to rise in LMICs (GSMA, 2020), and as programs increasingly rely on mobile phones and money to distribute benefits (cf. Gentilini et al., 2020), CDR-based targeting methods will likely play a more prominent role in the set of options considered by policymakers and program administrators — particularly in contexts like Afghanistan, where traditional targeting benchmarks are missing or unreliable. In just the past few years, for instance, data from mobile phone operators was used in the design of social assistance programs in Colombia, the Democratic Republic of the Congo, Pakistan, and Togo (Gentilini et al., 2020, 2021; Aiken et al., 2022). We conclude by highlighting a few policy considerations important for CDR-based targeting.

Speed and cost. An advantage of CDR-based targeting is that it can be used in contexts where face-to-face contact is not feasible, dramatically reducing the time required to implement a targeted program. While it typically takes many months (or years) to implement a proxy-means test (PMT), community-based targeting (CBT), or consumption-based targeting, a CDR-based model can be trained in just a few weeks (see Appendix C). Likewise, the marginal costs per household screened are substantially lower with CDR-based targeting than with CBT, PMT, or consumption-based targeting. For instance, Table S9 uses cost estimates obtained from the literature (and detailed in Table S10) to estimate targeting costs for the TUP program.[12] Whereas the marginal costs of screening an individual with a CBT or PMT are estimated at $2.20 and $4.00, respectively, the marginal cost of screening with CDR is negligible (see Appendix C).[13] For the entire TUP program, which screened around 125,721 households in six provinces, CBT and PMT would add an additional estimated $276,586 and $502,884, respectively, corresponding to 2.18% and 3.97% of the total program budget.

[10] A key nuance in this analysis is that for the CDR-based and combined methods where households without phones are targeted first, the precision and recall measures in Table 2 correspond to programs that only target households without phones (at random), as the number of households without phones exceeds the budget constraint of the program. The AUC score, on the other hand, is a summary statistic that represents targeting accuracy at all counterfactual targeting thresholds, and thus is not sensitive to the budget constraint — which explains the contrast between AUC and precision and recall in Table 2 Panels B and C. The ROC curves (Fig. 1) and Precision–Recall curves (Figure S5) highlight how budget constraint affects precision and recall.

[11] Due to the design of the TUP survey, which interviewed women in the household, we cannot avoid this mismatch between the survey respondent and the phone owner.

[12] In our cost calculations we obtain estimates for a CBT, rather than the hybrid approach used in the TUP program, as there is more information available on CBT-only costs in the literature. However, as the CBT cost can be interpreted as a lower bound for the cost of a hybrid approach, our qualitative results also apply to a hybrid approach.

[13] Marginal costs of CDR-based targeting are negligible because we assume no contact with screened individuals is required. In practice, it may be desirable to solicit informed consent to access CDR. If consent were collected in-person, the marginal costs would approach that of a PMT; if collected over the phone, there would still be significant cost savings, see Appendix Section C.

Data access and privacy. Access to phone data is necessary for CDR-based targeting. As we show, targeting performance degrades considerably when CDR are not available for subsets of the population. Encouragingly, the past several years have been characterized by a trend towards public sector access to CDR, particularly in the context of the COVID-19 pandemic, during which mobile network operators shared CDR with governments, researchers, and NGOs for social protection purposes (cf. Gentilini et al., 2020, 2021; Aiken et al., 2022). CDR have also been shared with the public sector for public health and humanitarian aid applications (Milusheva et al., 2021). Access issues aside, CDR contain private and sensitive data, including phone numbers and location traces. While much has been written about enabling responsible use of CDR for humanitarian response (e.g. de Montjoye et al., 2018; Oliver et al., 2020), to date no consistent privacy standards exist. Informed consent can increase participant agency, but also complicates the implementation logistics. Data minimization may provide a complementary pathway to privacy: as reflected in the feature importances in Table S2, our models rely primarily on only a fraction of the features we derive from mobile phone data — it may therefore be possible to restrict models to features that minimize privacy risk (such as statistics that do not involve contact networks or mobility patterns) without compromising model accuracy. Finally, there may be ways to incorporate differential privacy or other privacy enhancing technologies into a CDR-based targeting system, but such privatization would likely decrease targeting accuracy (Hu et al., 2015).

Algorithmic transparency and strategic behavior. Using CDR to determine program eligibility may introduce incentives for people to manipulate if and how they use their phone. For instance, while a program that targeted households without phones first might make sense in the context of one-off emergency response, it could not be deployed in equilibrium, as it would introduce undesirable incentives for people to not use their phones. In less extreme settings, we might still expect strategic manipulation of how people use their phones, if they know such behavior is being monitored. These considerations are not unique to CDR, as degrees of manipulation have been documented in social programs that use proxy means tests and other traditional targeting mechanisms (Camacho and Conover, 2011; Banerjee et al., 2018). While complex machine learning algorithms like the one presented in this paper may obfuscate the logic behind targeting decisions and thus reduce the scope for manipulation, this is not a 'solution.' Society often demands transparency in algorithmic decision-making, as black-box decisions are difficult to audit or hold to account. There is therefore a tension between the goals of increasing transparency and reducing manipulation, though recent advances in machine learning explore mechanisms for pursuing both objectives at once (Björkegren et al., 2020).

Centralized vs. local knowledge. CDR-based methods enable a top-down, centralized and standardized approach to program targeting, rather than a bottom-up approach that prioritizes local knowledge that can be elicited, for example, through community wealth rankings. While the empirical results in this paper indicate that the efficiency gains from CDR-based targeting are substantial, it may reinforce existing power structures (Taylor, 2016; Blumenstock, 2018a; Abebe et al., 2021). Efficiency gains should also be considered within the context of evidence suggesting that participating communities may prefer community-based approaches (Alatas et al., 2012), but also may perceive them as less legitimate (Premand and Schnitzer, 2020).

To summarize, our results suggest that there is potential for using CDR-based methods to determine eligibility for economic aid or interventions, substantially reducing program targeting overhead and costs. Our results also indicate that CDR-based methods may complement and enhance existing survey-based methods. We note, however, that the practical and ethical limitations to CDR-based targeting are significant. We emphasize the need to consider these limitations and the constraints of specific local contexts alongside the efficiency gains offered by CDR-based targeting.

Data availability

The authors do not have permission to share data.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jdeveco.2022.103016.

References

Abebe, R., Aruleba, K., Birhane, A., Kingsley, S., Obaido, G., Remy, S.L., Sadagopan, S., 2021. Narratives and Counternarratives on Data Sharing in Africa. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21, Association for Computing Machinery, New York, NY, USA, pp. 329–341. http://dx.doi.org/10.1145/3442188.3445897.

Aiken, E., Bellue, S., Karlan, D., Udry, C., Blumenstock, J.E., 2022. Machine learning and phone data can improve targeting of humanitarian aid. Nature 603 (7903), 864–870. http://dx.doi.org/10.1038/s41586-022-04484-9, https://www.nature.com/articles/s41586-022-04484-9, Number: 7903 Publisher: Nature Publishing Group.

Alatas, V., Banerjee, A., Hanna, R., Olken, B., Tobias, J., 2012. Targeting the poor: Evidence from a field experiment in Indonesia. Amer. Econ. Rev. 102 (4), 1206–1240.

Ali, S., May, M., 2021. Bangladesh's COVID-19 response is taking digital finance to new levels. https://www.cgap.org/blog/bangladeshs-covid-19-response-taking-digital-finance-new-levels.

Alkire, S., Foster, J., Seth, S., Santos, M.E., Roche, J.M., Ballon, P., 2015. Multidimensional Poverty Measurement and Analysis. Oxford University Press, URL https://EconPapers.repec.org/RePEc:oxp:obooks:9780199689491.

Banerjee, A., Duflo, E., Chattopadhyay, R., Shapiro, J., 2007. Targeting Efficiency: How Well Can We Identify the Poor? Working Paper Series No. 21, Institute for Financial Management and Research Centre for Micro Finance.

Banerjee, A., Hanna, R., Olken, B.A., Sumarto, S., 2018. The (lack of) Distortionary Effects of Proxy-Means Tests: Results from a Nationwide Experiment in Indonesia. Working Paper No. 25362, National Bureau of Economic Research, http://dx.doi.org/10.3386/w25362, URL http://www.nber.org/papers/w25362.

Bedoya, G., Coville, A., Haushofer, J., Isaqzadeh, M., Shapiro, J., 2019. No Household Left Behind: Afghanistan Targeting the Ultra Poor Impact Evaluation. World Bank Policy Research Working Paper 8877.

Björkegren, D., Blumenstock, J.E., Knight, S., 2020. Manipulation-proof machine learning. arXiv preprint arXiv:2004.03865.

Blumenstock, J., 2016. Fighting poverty with data. Science 353, 753–754.

Blumenstock, J., 2018a. Don't forget people in the use of big data for development. Nature 561, 170–172.

Blumenstock, J., 2018b. Estimating economic characteristics with phone data. Am. Econ. Rev.: Pap. Proc. 108, 72–76.

Blumenstock, J., 2020. Machine learning can help get COVID-19 aid to those who need it most. Nature http://dx.doi.org/10.1038/d41586-020-01393-7, URL https://www.nature.com/articles/d41586-020-01393-7.

Blumenstock, J., Cadamuro, G., On, R., 2015. Predicting poverty and wealth from mobile phone data. Science 350, 1073–1076.

Brown, C., Ravallion, M., van de Walle, D., 2018. A poor means test? Econometric targeting in Africa. J. Dev. Econ. 134, 109–124.

Burke, M., Driscoll, A., Lobell, D.B., Ermon, S., 2021. Using satellite imagery to understand and promote sustainable development. Science 371 (6535).

Camacho, A., Conover, E., 2011. Manipulation of Social Program Eligibility. Am. Econ. J.: Econ. Policy 3 (2), 41–65. http://dx.doi.org/10.1257/pol.3.2.41, URL https://www.aeaweb.org/articles?id=10.1257/pol.3.2.41.

Chi, G., Fang, H., Chatterjee, S., Blumenstock, J.E., 2022. Microestimates of wealth for all low- and middle-income countries. Proc. Natl. Acad. Sci. 119 (3), http://dx.doi.org/10.1073/pnas.2113658119, URL https://www.pnas.org/content/119/3/e2113658119, ISBN: 9782113658118 Publisher: National Academy of Sciences Section: Social Sciences.

Coady, D., Grosh, M., Hoddinott, J., 2004. Targeting outcomes redux. World Bank Res. Observer 19 (1).

Corral, P., Irwin, A., Krishnan, N., Mahler, D.G., 2020. Fragility and Conflict: on the Front Lines of the Fight Against Poverty. World Bank Publications.

de Montjoye, Y., Gambs, S., Blondel, V., Canright, G., De Cordes, N., Deletaille, S., Engø Monsen, K., Garcia-Herranz, M., Kendall, J., Kerry, C., et al., 2018. On the privacy-conscientious use of mobile phone data. Sci. Data 5 (1), 1–6.

de Montjoye, Y., Rocher, L., Pentland, A., 2016. Bandicoot: A Python toolbox for mobile phone metadata. J. Mach. Learn. Res. 17, 1–5.

Deaton, A., 1997. The Analysis of Household Surveys: a Microeconometric Approach to Development Policy. World Bank Publications.

Deaton, A., 2016. Measuring and understanding behavior, welfare, and poverty. Amer. Econ. Rev. 106 (6), 1221–1243. http://dx.doi.org/10.1257/aer.106.6.1221, URL https://www.aeaweb.org/articles?id=10.1257/aer.106.6.1221.

Engstrom, R., Hersh, J.S., Newhouse, D.L., 2017. Poverty from Space: Using High-Resolution Satellite Imagery for Estimating Economic Well-Being.

Fatehkia, M., Tingzon, I., Orden, A., Sy, S., Sekara, V., Garcia-Herranz, M., Weber, I., 2020. Mapping socioeconomic indicators using social media advertising data. EPJ Data Sci. 9 (1), 22.

Filmer, D., Pritchett, L., 2001. Wealth effects without expenditure data—or tears: An application to educational enrollments in states of India. Demography 39, 115–132.

Fortin, S., Kameli, Y., Kone, K., Belem, B., Sangho, H., Savy, M., 2018. Targeting vulnerable households in rural mali: Effectiveness of a community-based methodology, with or without addition of a proxy-mean test, 2016. Revue D'épidémiologie Et De Santé Publique 66, S353. http://dx.doi.org/10.1016/j.respe.2018.05.317, URL https://www.sciencedirect.com/science/article/pii/S0398762018310174, European Congress of Epidemiology "Crises, epidemiological transitions and the role of epidemiologists".

Gentilini, U., Almenfi, M., Orton, I., Dale, P., 2020. Social Protection and Jobs Responses to COVID-19: A Real-Time Review of Country Measures. World Bank Policy Brief, URL https://openknowledge.worldbank.org/handle/10986/33635.

Gentilini, U., Khosla, S., Almenfi, M., 2021. Cash in the City. World Bank, Washington, DC.

Grosh, M., Baker, J.L., 1995. Proxy Means Tests for Targeting Social Programs. Living Standards Measurement Study Working Paper 118, pp. 1–49.

Grosh, M., Leite, P., Wai-Poi, M., 2022. A New Look at Old Dilemmas: Revisiting Targeting in Social Assistance. World Bank Publications. The World Bank.

GSMA, 2020. Mobile economy. https://www.gsma.com/mobileeconomy/wp-content/uploads/2020/03/GSMA_MobileEconomy2020_Global.pdf.

Hanna, R., Olken, B., 2018. Universal basic incomes versus targeted transfers: Anti-poverty programs in developing countries. J. Econ. Perspect. 32, 201–226.

Hernandez, M., Hong, L., Frias-Martinez, V., Frias-Martinez, E., 2017. Estimating Poverty Using Cell Phone Data: Evidence from Guatemala. World Bank Policy Research Working Paper Series No. 7969.

Hu, X., Yuan, M., Yao, J., Deng, Y., Chen, L., Yang, Q., Guan, H., Zeng, J., 2015. Differential privacy in telco big data platform. Proc. VLDB Endow. 8 (12), 1692–1703.

Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., Ermon, S., 2016. Combining satellite imagery and machine learning to predict poverty. Science 353 (6301), 790–794. http://dx.doi.org/10.1126/science.aaf7894, URL http://science.sciencemag.org/content/353/6301/790.

Jerven, P., 2013. Poor Numbers. Cornell University Press.

Karlan, D., Lowe, M., Osei, R., Osei-Akoto, I., Roth, B., Udry, C., 2021. Cash transfers as COVID-19 relief: Evidence from Ghana. https://www.theigc.org/blog/cash-transfers-as-covid-19-relief-evidence-from-ghana/.

Karlan, D., Thuysbaert, B., 2019. Targeting ultra-poor households in Honduras and Peru. World Bank Econ. Rev. 33 (1), 63–94.

Lindert, K., Karippacheril, T.G., Caillava, I.R., Chávez, K.N., 2020. Sourcebook on the Foundations of Social Protection Delivery Systems. World Bank Publications.

Milusheva, S., Lewin, A., Gomez, T.B., Matekenya, D., Reid, K., 2021. Challenges and opportunities in accessing mobile phone data for COVID-19 response in developing countries. Data Policy 3.

Oliver, N., Lepri, B., Sterly, H., Lambiotte, R., Deletaille, S., De Nadai, M., Letouzé, E., Salah, A.A., Benjamins, R., Cattuto, C., et al., 2020. Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. Sci. Adv. 6 (23), eabc0764.

Paul, B.V., Msowoya, C., Archibald, E., Sichinga, M., Peredo, A.C., Malik, M.A.A., 2021. Malawi COVID-19 Urban Cash Intervention Process Evaluation Report. World Bank Publications.

Pokhriyal, N., Jacques, D., 2017. Combining disparate data sources for improved poverty prediction and mapping. Proc. Natl. Acad. Sci. 114, E9783–E9792.

Premand, P., Schnitzer, P., 2020. Efficiency, legitimacy, and impacts of targeting methods: Evidence from an experiment in Niger. World Bank Econ. Rev. http://dx.doi.org/10.1093/wber/lhaa019, lhaa019.

Ravallion, M., 1998. Poverty Lines in Theory and Practice, Vol. 133. World Bank Publications.

Schnitzer, P., Stoeffler, Q., 2021. Targeting for Social Safety Nets. World Bank, Washington, DC.

Sen, A., 1992. The Political Economy of Targeting. World Bank Washington, DC.

Sheehan, E., Meng, C., Tan, M., Uzkent, B., Jean, N., Lobell, D., Burke, M., Ermon, S., 2019. Predicting economic development using geolocated Wikipedia articles. In: Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.

Steele, J., Sundsø y, P., Pezzulo, C., Alegana, V., Bird, T., Blumenstock, J., Bjelland, J., Engø-Monsen, K., de Montjoye, Y., Iqbal, A., Hadiuzzaman, K., Lu, X., Wetter, E., Tatem, A., Bengtsson, L., 2017. Mapping poverty using mobile phone and satellite data. J. R. Soc. Interface 14.

Taylor, L., 2016. No place to hide? The ethics and analytics of tracking mobility using mobile phone data. Environ. Plann. D 34 (2), 319–336.

USAID, 2021. Usaid's direct cash transfer program helps over 85,000 vulnerable liberians cope with economic fallout from COVID-19. http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm.