Moving targets: When does a poverty prediction model need to be updated?*

Emily Aiken[†, §] U.C. Berkeley

Tim Ohlenburg[‡, §] University College London

Joshua Blumenstock[†] U.C. Berkeley

September 8, 2023

PRELIMINARY: DO NOT CITE OR CIRCULATE WITHOUT PERMISSION

Abstract

A key challenge in the design of effective social protection programs is determining who should be eligible for program benefits. In low and middle-income countries, one of the most common criteria is a Proxy Means Test (PMT) – a rudimentary application of machine learning that uses a short list of household characteristics to *predict* whether each household is poor, and therefore eligible, or non-poor, and therefore ineligible. Using nationwide survey data from six low and middle-income countries, this paper documents an important weakness in this use of machine learning: that the accuracy of the PMT prediction algorithm decreases steadily over time, by roughly 1.7 percentage points per year. We illustrate the implications of this finding for real-world anti-poverty programs, which typically update the PMT model only every 5-8 years, and then show that the aggregate effect can be decomposed into two forces: "model decay" caused by model drift, and "data decay" caused by changing household characteristics. Our final set of results show how an understanding of these forces can be used to optimize data collection policies to improve the effectiveness of social protection programs.

Keywords: keywords: proxy means tests, machine learning, poverty dynamics

*We thank Luis Inaki Alberro Encinas and Tina George for early input and inspiration for this project. We are grateful for financial support from Schmidt Futures and the National Science Foundation. Aiken gratefully acknowledges support from a Microsoft Research PhD Fellowship.

[†]University of California, Berkeley: emilyaiken@berkeley.edu and jblumenstock@berkeley.edu

[‡]University College London: email@timohlenburg.com

[§]These authors contributed equally to this work.

1 Introduction

Each year, over a trillion dollars are spent on social protection programs globally, making up on average 13% of each country's gross domestic product (ILO, 2021). To ensure that social protection benefits are prioritized to the poorest, many of these programs are *targeted*, providing benefits to only eligible households. Eligibility for program benefits in high-income countries is typically determined using administrative data on income from tax authorities in a practice called *means testing*. In low- and middle-income countries (LMICs), however, high-quality income data is typically unavailable, incomplete, or out-of-date (Jerven, 2013), rendering means testing infeasible.

Instead, an increasing number of LMIC social protection programs are targeted using *proxy means testing*. Proxy-means tests (PMTs) (Grosh and Baker, 1995) use a machine learning (ML) model to predict per capita household consumption from information on household assets and demographics. The ML model is trained on a sample household survey containing consumption expenditure labels; the model is then used to estimate the per capita consumption of every household for which asset and demographic data is available. Eligibility for social protection programs is determined based on these estimated consumption values in relation to a threshold below which households receive benefits. PMTs thus rely on (1) a *program registry* or *social registry* that contains information on roughly 10-50 characteristics of the household (such as demographic composition and household assets) for all potentially eligible households, and (2) a sample survey of a representative subset of households (usually roughly 5,000-20,000 households) for which detailed consumption expenditure data are collected along with the social registry variables. In settings with limited administrative capacity, social registry data are typically collected in a "PMT sweep" of households in parts or all of a country.

PMTs are now used in over fifty LMICs collectively containing over a billion people (Barrientos, 2018), making these decision rules one of the more widespread and consequential use cases of machine learning in government policy. It is therefore unsurprising that the performance of PMT-based eligibility has been widely studied in the economics literature (Hanna and Olken, 2018; Alatas et al., 2012; Premand and Schnitzer, 2021; Grosh and Baker, 1995; Noriega-Campero et al., 2020; McBride and Nichols, 2018; Hillebrecht et al., 2023), in contexts ranging from Indonesia (Alatas et al., 2012) to Niger (Premand and Schnitzer, 2021) to Peru (Hanna and Olken, 2018). Existing evaluations typically quantify the *exclusion errors* (the share of poor or beneficiary households incorrectly identified as non-poor by the PMT) and *inclusion errors* (the share of beneficiary households incorrectly

identified as eligible) resulting from PMT-based eligibility rules. In general, these evaluations find that while PMTs are imperfect, sometimes producing substantial errors of inclusion and exclusion, they often perform better than the other targeting approaches that would be feasible to implement.

However, the vast majority of this literature looks at the performance of a PMT at a single point in time — the moment when the PMT data are collected and the PMT decision rule is implemented — which is also when the performance of the PMT is highest. In practice, most PMT-based poverty registries are updated infrequently: while many social protection administrations aspire to update the social registry and ML model regularly (e.g. every two years in Costa Rica; every three years in Colombia, Indonesia, and Mexico (Barca and Hebbar, 2020)), in reality updates typically occur roughly every 5-8 years (Barca and Hebbar, 2020; Irarrázaval et al., 2011). This delay occurs because the updates are costly, involving some combination of (i) updating the household-level data in the registry with a "PMT sweep", which we estimate costs around $31 million in the median country (of the 15 for which we have data); and (ii) recalibrating the ML model with a new sample survey, which has a median cost of $1.3 million.

In the time between when the PMT data are collected and when policy decisions are made based on those data, the living conditions and poverty status of households may change. The rich literature on poverty dynamics has documented that households move in and out poverty (as determined by consumption expenditures) on a fairly regular basis: for example, Baulch and Hoddinott (2000) find that across panel studies in eight countries, the poverty status of 20-66% of households changes between survey waves. The PMT covariates collected in a social registry are likely to also shift over time: Kidd et al. (2021) show that at least one social registry covariate changes for 99.9% of households over a four year time span in Rwanda. Household size alone (a standard social registry covariate) changes for 87% of households.[1]

The machine learning literature has formalized these issues of temporal instability in more general time series machine learning applications as *dataset shift* (Quinonero-Candela et al., 2008). In this paper, we adapt the dataset shift framework from the machine learning literature to conceptualize and quantify how PMT accuracy is impacted by gaps in time between data collection and PMT deployment. Leveraging 25 rounds of survey data from

[1]Temporal stability of covariates has for some time been identified as a challenge for PMTs (Coady et al., 2004; Kidd and Wylde, 2011; Barca and Hebbar, 2020); while some PMTs have been designed to be in principle more robust to temporal instability by selecting only time-robust covariates (e.g. Tabor, 2002; Emmerling, 2012), to date these approaches have been ad-hoc.

six countries over twelve years, this paper makes three main contributions.

First, we quantify the total effect of allowing a PMT to become out-of-date. We find that each year that a PMT is not updated, it explains roughly 9 percentage points less of the variation in household consumption, resulting in an increase of inclusion and exclusion errors of 1.7 percentage points per year for a 30% coverage program (relative to an average up-to-date accuracy of 37.1% inclusion and exclusion errors).

Second, we decompose this overall decay — which assumes that neither the PMT model nor the underlying social registry data are updated — into *data decay* and *model decay*. Data decay is defined as losses in PMT accuracy due to infrequent collection of social registry data. As household conditions change, housing and demographic variables collected in the social registry are likely to become out-of-date, resulting in a *covariate shift* that is not corrected until a new PMT sweep is conducted. Model decay is defined as losses in PMT accuracy due to infrequent recalibration of the ML model, resulting in *model drift* in the learned relationship between the social registry data and consumption expenditure.[2] Our results identify that data decay is approximately three times more powerful than model decay in reducing PMT accuracy.

Third, and finally, we use international information on survey costs to assess the financial implications of data and model recalibration policies available to social program administrators. We find that, under reasonable assumptions about the trade-off between survey costs and the cost of mis-targeted benefits, most social protection programs should aim to collect social registry data and recalibrate the PMT model every 1-3 years.

This paper provides the first comprehensive empirical assessment of the impacts of model and data decay on PMT accuracy. Most published PMT evaluations do not account for either data decay or model decay over time, relying on social registry and sample survey data collected in a single, simultaneous effort to assess targeting accuracy. A handful of previous papers have taken limited steps towards measuring one or the other of model decay or data decay (see Table 1). However, existing work is not standardized across multiple countries, does not separately consider model and data decay (and the implications for policy of these two different forces), and does not look at effects after three years (when in practice most programs face delays of 5-8 years (Barca and Hebbar, 2020)). By providing cross-country evidence on the returns to data updating for PMT accuracy, this paper provides both a

[2]Model drift has been widely studied in the machine learning literature (Sugiyama and Kawanabe, 2012; Zhang et al., 2013; Koh et al., 2021), and in real-world ML applications where the underlying joint distribution of input variables and the predictive target may change over time, ranging from medical diagnosis (Davis et al., 2017) to forecasting market volatility (Gibbs and Candes, 2021) and classifying media articles (Yao et al., 2022)

general assessment of accuracy decay over time and a framework that policymakers can adapt to identify decay effects and assess updating policies in specific country contexts.

2 Methods

2.1 Data

Our analysis relies on publicly available panel surveys from the Living Standards Measurement Study (LSMS) in Ethiopia, Nigeria, Tanzania, and Uganda[3], the Ghana Panel Survey[4], and Peru's Encuesta National de Hogares.[5] Each panel is between three and five rounds, covering between five and eleven years (Table 2). All surveys were conducted between 2008 and 2019. Since our analysis relies on observing changing households conditions, we consider only households that appear in all rounds of the survey in each country. The resulting sample sizes range from 424 households in Tanzania to 3,393 households in Ghana.

2.2 Proxy means test modelling

Following the standard implementation of a PMT (Grosh and Baker, 1995; Brown et al., 2018; McBride and Nichols, 2018), we construct a machine learning experiment in which log-transformed per capita household consumption expenditure is estimated based on housing-related and demographic covariates collected in the survey.[6] We refer to the predictor variables collectively as the "registry covariates" as they are the types of variables that would typically be collected in a social or program registry (Grosh and Baker, 1995; Brown et al., 2018; Hanna and Olken, 2018). We use three types of registry covarites in our models: (1) housing-related variables (such as the material of the roof, walls, and floor; amenities such as a toilet or electricity; and information on building size and ownership), (2) asset ownership variables (unique to each country's context), and (3) demographic information (such as household size, number of children, and characteristics of the household head).

[3]Information and microdata are available from `https://www.worldbank.org/en/programs/lsms/initiatives/lsms-ISA`. For each LSMS survey, we use all survey waves for which consumption panel data are available.

[4]Information and microdata are available from `https://egc.yale.edu/data/isser-northwestern-yale-long-term-ghana-socioeconomic-panel-survey-gsps`. We use all three survey waves.

[5]Information and microdata are available from `https://iinei.inei.gob.pe/microdatos/`. We use the latest consecutive panel survey rounds prior to the COVID-19 pandemic, 2015-2019.

[6]Some PMT implementations use per-capita consumption values as the predictive target while others use adult equivalence weighting; for consistency we use per-capita values across countries.

Not all covariates are available in each of the six countries we study, so the number of covariates ranges from 28 in Peru to 68 in Tanzania. Continuous variables are winsorized with a 99% limit and are then standardized to a [0, 1] range and categorical covariates are one-hot encoded.[7]

2.2.1 Machine learning approach

Our base specification divides surveyed households for each country randomly into a training set (75% of households) and an evaluation set (25% of households). For each survey round separately, we train our machine learning models to predict log-transformed per capita household consumption from the social registry covariates on training set households, and evaluate performance on test set households. Survey weights are included both in training the ML models and in calculating the accuracy metrics. To account for idiosyncrasies in random data splits, we repeat the simulation 100 times with different random train-test splits and take the average of each evaluation metric across test sets.

The primary machine learning approach we test is a linear regression paired with stepwise forward selection of input variables. Although there are a variety of machine learning models used for PMTs in practice, and recent work has suggested that more complex models may provide marginal accuracy gains (McBride and Nichols, 2018; Noriega-Campero et al., 2020; Areias and Wai-Poi, 2022), the stepwise forward selection and linear regression approach is a standard approach to calibrating PMTs.[8] We also compare the performance of the stepwise approach with a number of other machine learning models, including ordinary least squares regression (with the full set of registry covariates), LASSO regression, a random forest, and a gradient boosting machine.[9]

[7]For categorical variables, any value representing less than 1% of the observations in any survey round is replaced with a generalized "other" category.

[8]Our implementation of stepwise forward selection is as follows, performed separately for each random train-evaluation data split: First, the training set is split further into a selection set and a calibration set. The linear regression is trained only on the selection set; variables are progressively added to the model until the mean squared error of the model on the calibration set stops increasing. The linear regression is then retrained on the entire training set prior to the calculation of performance metrics on the evaluation set.

[9]For the LASSO regression, regularization strength chosen via three-fold cross validation on the training set. For the random forest, ensemble size and maximum depth chosen via three-fold cross validation. For the gradient boosting machine, the ensemble size, maximum depth, minimum samples per leaf, and learning rate chosen via three-fold cross validation.

2.2.2 Evaluation metrics

To evaluate PMT performance, we calculate the R^2 score and Spearman's rank correlation between predicted and ground truth consumption values using the evaluation set. We also calculate the targeting error rate of the PMT for a hypothetical social protection programs that aim to target the poorest 10%, 20%, 30%, and 40% of households. We use a *quota approach* to calculating targeting error rates: if a household is ranked in the poorest 10% by ground-truth consumption but not by predicted consumption it is an error of exclusion; if a household is ranked in the poorest 10% by predicted consumption but not by ground truth consumption it is an error of inclusion. In this setting the exclusion error rate and inclusion rate are equal by definition; following Brown et al. (2018), we therefore refer to this metric as the *targeting error rate (TER)*. We notate the targeting error rate with the relevant quota: for example, the targeting error rate for a program aiming to reach the poorest 20% of households is notated *TER(20)*. All evaluation metrics are calculated separately for the 100 simulations with different random data splits, and the average across the splits is reported.

2.3 Quantifying decay

2.3.1 Combined decay

To simulate decay caused by temporal gaps between data collection, model calibration, and deployment, we conduct the same machine learning experiment described in the base specification above, but this time introduce lags between training and evaluation sets. We start by estimating combined decay: the setting in which the model and social registry covariates are equally out-of-date. To simulate combined decay, the ML model is trained on the training set from round w_i, and performance metrics are calculated on the evaluation set using covariates from round w_i and consumption data from round w_j, for all $i <= j$ in the panel survey. Across the six countries and all possible lag values, we have 67 observations of combined decay, ranging from lags of 1 year to lags of 10 years. To parameterize combined decay, we regress decay (that is, the difference between performance with temporal lags and performance using data that matches the time of the evaluation) on the temporal lapse:

$$\Delta accuracy_t^{k,l} = \beta_{combined} t + \epsilon_t^{k,l} \tag{1}$$

$\beta_{combined}$ is the parameter of interest, representing the decay in PMT accuracy associated with each year that social registry data are not updated. $\Delta accuracy_t^{k,l} = accuracy_0^{k,l} - accuracy_t^{k,l}$, where t indexes the years since data collection and model recalibration, and k

indexes each country, and l indexes the year of test data collection. The specification does not include an intercept, as decay is zero when both the model and data are up-to-date.

2.3.2 Decomposition into model and data decay

To estimate the contributions of model and data decay to combined decay in PMT accuracy, we expand our dataset to include lag combinations where the model and data are unequally out-of-date. In this setting, the ML model is trained on the training set from round w_i, and performance metrics are calculated on the evaluation set using covariates from round w_j and consumption data from round w_k, for all $i, j <= k$ in the panel survey, for a total of 223 observations of decay across the six panel surveys. In our decomposed specification, we regress decay on the time since model calibration, the time since data collection, and the interaction of the two:

$$\Delta accuracy_{u,v}^{k,l} = \beta_{data} u + \beta_{model} v + \beta_{interaction} uv + \epsilon_{u,v}^{k,l} \tag{2}$$

The inclusion of the interaction term $\beta_{interaction}$ accounts for the way in which model and data decay co-vary, and allows a more flexible fit as the lags increase. In the decomposed specification $\Delta accuracy_{t,s}^{k,l} = accuracy_{0,0}^{k,l} - accuracy_{u,v}^{k,l}$, where u indexes the years since PMT data collection, v indicates the years since model recalibration, k indexes each country, and l indexes the year of test data collection.

2.4 Choice of data collection policy

We consider periodic data updating policies that prescribe one fixed annual interval for ML model recalibration and another, possibly identical one, for PMT sweeps. In this setting, a policymaker wishing to minimise the total cost arising from accuracy decay, including survey expenses, faces the following choice of data collection policy:

$$\arg\min_{a,b} \frac{1}{T} \sum_{t=1}^{T} \Delta TER_{i,j} * \pi^t + \mathbb{1}(i=0) c_{SPS}^t + \mathbb{1}(j=0) c_{sweep}^t \forall a, b \in \{1..10\} \tag{3}$$

where a and b are the respective annual intervals at which the population sample survey and the PMT sweep are conducted, ΔTER is the combined decay of the targeting error rate, the evaluation period $T = a * b$ ensures that the lifecycle of the periodic policy, i.e. the number of years until both surveys are recollected simultaneously, is completed, $i = \text{mod}(a, t)$ and $j = \text{mod}(b, t)$ are modulo operations that yield the years since the sample

survey and household characteristics were collected, respectively, $\mathbb{1}()$ is an indicator function that equals one if the term in brackets is true and zero otherwise, π^t is the average benefit paid to households in the social registry in year t, and c^t_{SPS}, c^t_{sweep} are the costs respective costs of a sample population survey and a PMT sweep in year t. For simplicity, we apply no discount factor, and assume that $c^r = c^s$ and $\pi^r = \pi^s \forall r, s \in \mathbb{Z}$ so that neither benefit amounts nor costs change over T.

In our simulation, we test the 100 model and data refresh policy combinations that arise from 1-10 year update cycles for each. The best such policy is one that minimizes the total cost of social registry data collection, consumption survey data collection for calibrating the PMT model, and benefit payments lost to mis-targeting, over the course of the policy's lifecycle. It will critically depend on the program's coverage, the size of the social registry, and the benefits the program provides; we combine these latter two into a single parameter: the average benefits per household in the social registry.

2.4.1 Cost of PMT updating

Our cost estimates for social registry data updating come from several papers that have reported the per-household cost of conducting a PMT survey; in total we identified reported PMT survey costs in eight countries from four papers (Alatas et al., 2012; Karlan and Thuysbaert, 2019; Rosas et al., 2016; Schnitzer and Stoeffler, 2022). To make costs consistent across contexts, we measure all costs in 2015 purchasing power parity (PPP) dollars. Reported PMT survey costs range from \$6.32 to \$29.39 PPP (\$2.62 to \$12.00 nominal) per household in the social registry (Table S1); in our analysis we use the median value of \$13.31 PPP.

For model recalibration, the dominant cost is conducting a consumption sample survey to train the PMT. We obtain data on the per-survey cost and total sample size for LSMS consumption sample surveys 18 countries from Kilic et al. (2017). Total costs for the consumption survey range from \$245,784 nominal for a 5,016-household consumption survey in Kyrgyzstan to \$4.29 million nominal for a 14,400 household survey in Yemen (Table S2). Like for PMT costs, we distribute the consumption survey cost over all households in the hypothetical social registry; to obtain this value we calculate an estimate of the total number of households in each of the 18 countries.[10] Using the median global social registry coverage of 21% calculated by Grosh et al. (2022) this yields the expected number households in a

[10]Estimated households per country are calculated using population data from the World Bank (World Bank, 2023) and average household size data from the Nations (2017).

hypothetical social registry of each country. When the cost of consumption surveys is distributed over the households in these hypothetical registries, the average cost ranges from \$0.40 to \$10.87 PPP; in our analysis we use the median value of \$3.33 PPP.

2.4.2 Cost of mistargeting

From the perspective of targeting for a specific social assistance program, the cost of mistargeting due to decay is the value of benefits provided to households that are *additional* errors of inclusion due model or data decay. We therefore calculate the cost of mistargeting due to PMT decay *per social registry household* as the product of the decay in the targeting error rate and the average benefits per household in the social registry. The decay in targeting error rate expected for each recalibration policy is calculated using our parametric estimate of decay from Equation 2; the decay for each updating strategy is the average decay over the policy's entire lifecycle.

We calculate the updating and mistargeting costs of the 10*10=100 policy options for four different program coverage rates: 10%, 20%, 30%, 40%. The average benefit amounts per social registry household we consider ranges from \$10 to \$2,000 PPP. Note that this latter value does not refer to the benefits provided to beneficiaries, but rather the average benefits provided to any household in the social registry (measured in PPP dollars).[11]

3 Results

3.1 Quantifying decay

We begin by quantifying what happens to the accuracy of a PMT if neither the the PMT model nor the underlying data are updated. The effect of this *combined decay* (so called because it is produced by the combination of the model and data growing out of date) is shown across all six countries in Figure 1. Over the ten-year period shown, we see substantial reduction in the R^2 and Spearman correlation coefficient — and substantial increases in the targeting error rate — in each of the countries. The black line in Figure 1 is the regression line (see Equation 1); the coefficients of this regression line are provided in Table 3.

The PMT's accuracy degrades significantly over time: for each year that a PMT is not updated, the PMT explains roughly six percentage points less of the variation in household consumption. Inclusion and exclusion errors increase by between 1.5 and 1.9 percentage

[11] We prefer this approach as it abstracts away the need to quantify the size of the social registry in our simulations.

points each year, depending on the total coverage of the program. To put these numbers in context, with a program the size of PROGRESA in Mexico (which provides benefits to approximately 2.6 million households annually (Skoufias, 2005)), a delay of five years between PMT updates would be expected to produce roughly an additional 200,000 exclusion and inclusion errors. In the discussion section, we revisit these and other implications in greater detail.

Figure 1 demonstrates that decay rates are different across countries; to assess cross-country variation Table S5 calculates combined decay separately in each country. For example, the yearly decline in R^2 varies from 3 percentage points in Peru to 14 percentage points in Uganda, and the yearly increase in inclusion and exclusion error rates (for a 30% coverage program) varies from 0.8 percentage points in Tanzania to 3.9 percentage points in Uganda.

3.1.1 Decomposition into model and data decay

Our next step is to identify how much of the combined decay can be attributed to *model decay* (model drift in the relationship between social registry covariates and ground truth consumption, which occurs because the model is not recalibrated with sample survey data) and *data decay* (covariate shift in the social registry, which occurs because the household covariates in the registry are not updated through PMT sweeps). To better understand the relative contributions of model and data decay separately, we assess how PMT accuracy varies with the years since model recalibration, the years since data collection, and the interaction of the two lags (see Equation 2). Table 4 presents the results of these regressions.

We find that both model decay and data decay significantly impact PMT accuracy, but that data decay has the largest impact on errors of inclusion and exclusion, contributing roughly three times as much to overall decay: the decay parameters are 1.85 percentage points per year and 0.65 percentage points per year for data decay and model decay respectively (for a program with 30% coverage, with similar effects for other program coverage levels). Decay contributions are similar for Spearman correlation (with a parameter of -0.0347 for data decay and -0.0115 for model decay), although model decay and data decay contribute approximately equally to decreases in the R^2 score (parameters of -0.0544 and -0.0479, respectively). The interaction term is positive and statistically significant for R^2 and Spearman, and negative and statistically significant for all targeting error rates. The interaction term therefore "works against" model decay and data decay, resulting in a flattening of overall decay at long joint data lags.

3.1.2 Other ML approaches

Past work has shown that more sophisticated machine learning models may produce more accurate PMT predictions than the standard OLS or OLS with stepwise forward selection modeling approach (McBride and Nichols, 2018; Noriega-Campero et al., 2020; Areias and Wai-Poi, 2022). In our setting, it is also possible that more sophisticated machine learning approaches could be more robust to model and data decay over time. To test this possibility, Table S3 measures the baseline (no-decay) performance of a number of ML models (OLS, OLS with stepwise forward selection, LASSO, a random forest, and a gradient boosting machine), and Table S4 measures overall decay for each of these models.

Consistent with past work we find that more complex and nonlinear ML models are slightly more accurate at baseline than an OLS or OLS plus stepwise forward selection approach. A gradient boosting model has the highest baseline accuracy — in terms of targeting error rate — of any of the ML approaches tested. In general, we find similar patterns of decay across ML models. Of all the models tested, the linear regression with stepwise forward selection and LASSO regression perform slightly better than any others in terms of slowing decay — but the differences are relatively small.

3.2 Assessing model and data refresh policies

Data collection for model re-calibration and PMT sweeps is costly, but in its absence accuracy decay leads to a mis-allocation of social assistance spending towards households that are ineligible. The resulting dilemma for policymakers is how to balance cost and decay economically.

We test a set of 100 model and data refresh policies, representing every combination of social registry data updating from every 1-10 years, and likewise for PMT model recalibration. As described in detail in Section 2.4, we consider the best updating policy as the one that minimizes the total cost of social registry data collection, consumption survey data collection for calibrating the PMT model, and benefit payments lost to mis-targeting, over the course of the policy's lifecycle. The best updating policy will critically depend on the program's coverage, the size of the social registry, and the benefits the program provides; we combine these latter two into a single parameter: the average benefits per household in the social registry.

Figure 2 shows the minimum-cost, periodic updating strategy (in terms of frequency of PMT sweeps to collect social registry data and frequency of model recalibration) for

programs targeting the poorest 10%, 20%, 30%, and 40% of households, and for average benefits per household in the social registry ranging from $0 to $2,000 PPP. We find that programs with moderately large benefit values per household in the social registry, starting from around $250 PPP per year, should adopt frequent data collection policies (recalibrating the model and conducting a PMT sweep every 1-2 years). Programs with very small benefit values (under $80 PPP per household in the social registry) should only update rarely, approximately every 6-10 years for both the model and the social registry data. Programs with intermediate payments, i.e. those providing $80-250 PPP per household in the social registry, should invest in updating the model and the social registry data every 2-4 years.

To put these results in the context of real-world programs, Table 5 provides details of thirteen real-world social protection systems for which data on design parameters (including social registry coverage and total program budget) are available from the Manchester Social Assistance Explorer (Barrientos, 2018). We include our own estimates of the optimal updating strategy based on each program's design and the results in Figure 2.[12] While information on the status quo updating policies for specific programs is not generally available, Barca and Hebbar (2020) estimate an average of 5-8 year gaps between PMT sweeps. In general, we find that the recommended updating policy is more frequent: most real-world programs should update the social registry data and the ML model every 1-3 years. The exact recommended updating strategy depends on the design parameters of the program.

In Appendix B we test an alternative approach to identifying the best updating strategy, by assuming a fixed and single budget for data collection and program benefits and optimizing for social welfare using a utility function. We find that this approach to selecting the best updating strategy tends even further towards prescribing frequent updating (for example, prescribing a yearly ML model update and survey sweep for all programs with benefits of over $60 PPP per social registry household, Figure S2).

4 Discussion

This article investigates the practical impact of a statistical phenomenon — dataset shift — through its influence on the widely applied predictive modelling task of proxy means testing. It suggests that around 1.7% of intended social assistance beneficiaries are excluded due to dataset shift if the ML model and registry covariates used in PMTs are allowed to go out-of-date by a single year. By year five, which may be a fair estimate of the lag between

[12] We assume a 30% program coverage level, although updating recommendations are fairly robust across alternative coverage levels.

data collection rounds (Barca and Hebbar, 2020; Irarrázaval et al., 2011), that number has risen to around 8.5% of beneficiaries. These figures and the scale of targeted social assistance programs (Barrientos, 2018; Gentilini et al., 2022) suggest that millions suffer the effects of accuracy decay each year.

Data decay has approximately three times the impact on overall performance of model decay. This result is in line with insights from studies of poverty dynamics showing that a substantial number of households suffer idiosyncratic shocks in a given year (Baulch and Hoddinott, 2000). The failure to capture this variation seems to weigh more heavily than use of an ageing predictive model. Our results also show a moderately negative interaction effect, signalling the importance of aligning the model with the period of covariate data collection, even when the covariates are out-of-date. The rates of model and data decay are surprisingly consistent across program coverage levels (in spite of large differences in baseline accuracy levels).

In comparison with previous studies on the topic, Brown et al. (2018) found a larger combined decay effect (7-9 percentage points per year), but the lack of out-of-sample validation in the non-decay scenario does not allow for direct comparison to this result. In contrast, Klasen and Lange (2015) and Sebastian et al. (2018) find little evidence of model or data decay; perhaps due to investigating a small effect with limited data over a short time horizon (1-3 years) for a single country (Bolivia and Sri Lanka, respectively). Our estimate is most consistent with that of Hillebrecht et al. (2023), which identifies a combined decay effect of 2 percentage points per year in Burkina Faso.

Based on our results, the recommended model recalibration policy for an average setting is every 1-2 years for programs with benefits over $250 PPP per social registry household. The fast amortization of consumption sample survey costs suggests that countries with meaningful household-targeted social assistance programs should collect consumption data to recalibrate the ML model annually or bi-annually. Non-targeting uses only the bolster the case for frequent socio-economic surveys, and this study also illustrates the benefit of panel data for social protection design.

Households sweeps are roughly four times as costly as sample surveys per social registry member, but their large accuracy impact implies that they should nevertheless be conducted at approxoimately the same frequency as model recalibration: every 1-2 years for programs providing benefits of over $250 PPP per social regsitry household. While this suggested updating frequency is for the most part consistent with program commitments and policy guidance on updating frequency (Barca and Hebbar, 2020; Irarrázaval et al., 2011; Sebastian

et al., 2018), it is substantially more frequent than the 5-8 year updating gaps that programs typically face in practice (Barca and Hebbar, 2020). Our empirical results suggest that policymakers should face the cost implications of accuracy decay by making contributions to a ring-fenced, accumulating survey fund through the annual budget of targeted social protection programs. Moreover, as a country's social protection system matures and average benefit amounts increase, more frequent data collection is advisable. Technological progress that lowers data collection costs, such as integration of digital administrative databases or the use of novel data sources, also calls for a higher frequency of model and household data updates.

In settings where the PMT data and/or model are very out-of-date and updating is not feasible, alternative targeting approaches might be preferable from an accuracy perspective until the PMT can be updated. Community-based targeting is generally found to be only slightly less accurate than the PMT, (Alatas et al., 2012; Karlan and Thuysbaert, 2019) and could be a suitable choice in settings where PMTs are out-of-date. Alternative digital data sources may also become relevant for targeting when PMTs become out-of-date: for example, Aiken et al. (2022) find a ten percentage point difference in targeting accuracy for the poorest 30% (TER(30)) between a freshly-calibrated PMT and targeting on a poverty index inferred from mobile phone metadata; our results suggest that if a PMT is six or more years out-of-date the phone-based approach may in fact be preferred to the PMT. Other simple targeting approaches — like geographic targeting (Baker and Grosh, 1994), categorical targeting (Devereux et al., 2017), or self-targeting (Alatas et al., 2016) may also be relevant policy alternatives when PMTs are very out-of-date.

The remarkable consistency of the decay results across program coverage levels is mirrored by surprisingly similar recommended updating policies for different coverage rates. Although this result suggests a certain robustness across settings, the average costs used in the policy simulations are not necessarily reflective of specific country conditions. Our framework can be adapted to calculate the best updating policy for a particular country given country-specific consumption survey and PMT survey costs to form more targeted estimates.

A few limitations to our analysis are worth noting. First, the decay estimates are derived from a limited set of six countries, with a geographic bias towards Sub-Saharan Africa. While the baseline targeting accuracies of the PMTs we simulate are broadly similar to other published work (Brown et al., 2018; Hanna and Olken, 2018; Alatas et al., 2012; Schnitzer and Stoeffler, 2022; Karlan and Thuysbaert, 2019), confirming that decay evolves in a similar fashion in other contexts would be helpful and policy-relevant. Second, we

use simple variable selection policies and limited feature engineering in the construction of predictive models, which are all trained on LSMS panel surveys of moderate size. Future work could confirm whether modelling and training data details matter, and whether more decay-robust ML models can be designed that also achieve high current accuracy. Third, while the periodic policies we consider are consistent with the framework of current updating strategies for most social protection programs (Barca and Hebbar, 2020), it is possible that more complex recalibration strategies may yield lower costs. Fourth, and finally, a further detail to consider in the simulations of updating policies is the time value of money and cost inflation, which we disregard for simplicity.

Our analysis also by design abstracts away certain aspects of real-world social protection systems, which may complicate the results presented here. First, we focus on settings where PMT sweeps are conducted to collect social registry data; the setting of on-demand registration common in social protection systems with elevated administrative capacity raises questions around comparability of rankings over time and suggests the potential for targeted reassessment. Second, there is additional loss in social registry accuracy and completeness resulting from the creation of new households over time (see Bah et al. (2019) on the importance of complete household lists); this further form of decay could be investigated with suitable panel data. Third, and finally, our results rely on sample survey data from the LSMS; further work could compare these simulation results with empirical data on real-world social protection program recipients over time.

In conclusion, this work establishes multi-country temporal accuracy decay estimates for PMTs and proposes initial practical cost-conscious mitigation policies. Our results suggest that accuracy decay over time is a first order concern for the targeting of social protection programs in LMICs, and that social registry data and PMT models should be updated substantially more frequently than the status quo. While accuracy decay is a serious concern for PMT models, it is unclear how alternative targeting methods fare in this regard. Alternative approaches that may be conducted frequently at low cost, such as some satellite-based geographical or phone-based targeting, are typically less accurate to start with; more research would be needed to establish whether they could perform better over time.

References

Aiken, E., Bellue, S., Karlan, D., Udry, C., and Blumenstock, J. E. (2022). Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, 603(7903):864–870.

Alatas, V., Banerjee, A., Hanna, R., Olken, B. A., and Tobias, J. (2012). Targeting the poor: evidence from a field experiment in indonesia. *American Economic Review*, 102(4):1206–1240.

Alatas, V., Purnamasari, R., Wai-Poi, M., Banerjee, A., Olken, B. A., and Hanna, R. (2016). Self-targeting: Evidence from a field experiment in indonesia. *Journal of Political Economy*, 124(2):371–427.

Areias, A. and Wai-Poi, M. (2022). Machine learning and prediction of beneficiary eligibility for social protection programs. *Revisiting Targeting in Social Assistance*, page 507.

Bah, A., Bazzi, S., Sumarto, S., and Tobias, J. (2019). Finding the poor vs. measuring their poverty: Exploring the drivers of targeting effectiveness in indonesia. *The World Bank Economic Review*, 33(3):573–597.

Baker, J. L. and Grosh, M. E. (1994). Poverty reduction through geographic targeting: How well does it work? *World development*, 22(7):983–995.

Barca, V. and Hebbar, M. (2020). On-demand and up to date? dynamic inclusion and data updating for social assistance. *GIZ (https://socialprotection.org/sites/default/files/publications_files/GIZ_DataUpdatingForSocialAssistance_3. pdf)*.

Barrientos, A. (2018). Social assistance in low and middle income countries 2000-2015.

Baulch, B. and Hoddinott, J. (2000). Economic mobility and poverty dynamics in developing countries. *The Journal of Development Studies*, 36(6):1–24.

Beegle, K., Coudouel, A., and Monsalve, E. (2018). *Realizing the full potential of social safety nets in Africa*. World Bank Publications.

Berner, H. and Van Hemelryck, T. (2021). Social information systems and registries of recipients of non-contributory social protection in latin america in response to covid-19.

Brown, C., Ravallion, M., and Van de Walle, D. (2018). A poor means test? econometric targeting in africa. *Journal of Development Economics*, 134:109–124.

Coady, D., Grosh, M., and Hoddinott, J. (2004). Targeting outcomes redux. *The World Bank Research Observer*, 19(1):61–85.

Davis, S. E., Lasko, T. A., Chen, G., Siew, E. D., and Matheny, M. E. (2017). Calibration drift in regression and machine learning models for acute kidney injury. *Journal of the American Medical Informatics Association*, 24(6):1052–1061.

Devereux, S., Masset, E., Sabates-Wheeler, R., Samson, M., Rivas, A.-M., and te Lintelo, D. (2017). The targeting effectiveness of social transfers. *Journal of Development Effectiveness*, 9(2):162–211.

Emmerling, J. (2012). Targeting cash transfers in mali–a proxy means test approach.

Gentilini, U., Almenfi, M. B. A., Iyengar, T., Okamura, Y., Downes, J. A., Dale, P., Weber, M., Newhouse, D. L., Rodriguez Alas, C. P., Kamran, M., et al. (2022). Social protection and jobs responses to covid-19.

Gibbs, I. and Candes, E. (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672.

Grosh, M. and Baker, J. L. (1995). Proxy means tests for targeting social programs. *Living standards measurement study working paper*, 118:1–49.

Grosh, M., Leite, P., Wai-Poi, M., and Tesliuc, E. (2022). *Revisiting targeting in social assistance: A new look at old dilemmas*. World Bank Publications.

Hanna, R. and Olken, B. A. (2018). Universal basic incomes versus targeted transfers: Anti-poverty programs in developing countries. *Journal of Economic Perspectives*, 32(4):201–26.

Hillebrecht, M., Klonner, S., and Pacere, N. A. (2023). The dynamics of poverty targeting. *Journal of Development Economics*, 161:103033.

ILO (2021). World social protection report 2020–22: Social protection at the crossroads–in pursuit of a better future.

IMF (2023). World economic outlook database. `https://www.imf.org/en/Publications/WEO/weo-database/2023/April/download-entire-database`.

Irarrázaval, I. et al. (2011). Sole information systems on beneficiaries in latin america. Technical report, Inter-American Development Bank.

Jerven, M. (2013). *Poor numbers: how we are misled by African development statistics and what to do about it*. Cornell University Press.

Karlan, D. and Thuysbaert, B. (2019). Targeting ultra-poor households in honduras and peru. *The World Bank Economic Review*, 33(1):63–94.

Kidd, S., Athias, D., and Mohamud, I. (2021). Social registries: A short history of abject failure. *Development Pathways*.

Kidd, S. and Wylde, E. (2011). Targeting the poorest: An assessment of the proxy means test methodology. *AusAID Research Paper, Australian Agency for International Development, Canberra, Australia*.

Kilic, T., Serajuddin, U., Uematsu, H., and Yoshida, N. (2017). Costing household surveys for monitoring progress toward ending extreme poverty and boosting shared prosperity. *World Bank Policy Research Working Paper*, (7951).

Klasen, S. and Lange, S. (2015). Targeting performance and poverty effects of proxy means-tested transfers: Trade-offs and challenges. Technical report, IAI Discussion Papers.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. (2021). Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR.

Leite, P., George, T., Sun, C., Jones, T., and Lindert, K. (2017). *Social registries for social assistance and beyond: a guidance note and assessment tool*. World Bank.

McBride, L. and Nichols, A. (2018). Retooling poverty targeting using out-of-sample validation and machine learning. *The World Bank Economic Review*, 32(3):531–550.

Nations, U. (2017). Household size and composition around the world. *Economic and Social Affairs*.

Noriega-Campero, A., Garcia-Bulle, B., Cantu, L. F., Bakker, M. A., Tejerina, L., and Pentland, A. (2020). Algorithmic targeting of social policies: fairness, accuracy, and distributed governance. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 241–251.

Ortiz, I., Duran, F., Pal, K., Behrendt, C., and Acuña-Ulate, A. (2017). Universal social protection floors: Costing estimates and affordability in 57 lower income countries. *ILO Extension of Social Security Working Paper*, (58).

Premand, P. and Schnitzer, P. (2021). Efficiency, legitimacy, and impacts of targeting methods: Evidence from an experiment in niger. *The World Bank Economic Review*, 35(4):892–920.

Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2008). *Dataset shift in machine learning*. Mit Press.

Rosas, N., Pinzón-Caicedo, M., and Zaldivar, S. (2016). Evaluating tanzania's productive social safety net: targeting performance, beneficiary profile, and other baseline findings. *Washington (DC): World Bank*.

Schnitzer, P. and Stoeffler, Q. (2022). Targeting for social safety nets: Evidence from nine programs in the sahel. *Available at SSRN 4017172*.

Sebastian, A. R., Shivakumaran, S., Silwal, A. R., Newhouse, D. L., Walker, T. F., and Yoshida, N. (2018). A proxy means test for sri lanka. *World Bank Policy Research Working Paper*, (8605).

Skoufias, E. (2005). *PROGRESA and its impacts on the welfare of rural households in Mexico*, volume 139. Intl Food Policy Res Inst.

Sugiyama, M. and Kawanabe, M. (2012). *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press.

Tabor, S. R. (2002). Assisting the poor with cash: Design and implementation of social transfer programs. *World Bank Social Protection Discussion Paper*, 223:79–97.

World Bank (2023). World bank open data. `https://data.worldbank.org/`.

Yao, H., Choi, C., Cao, B., Lee, Y., Koh, P. W. W., and Finn, C. (2022). Wild-time: A benchmark of in-the-wild distribution shift over time. *Advances in Neural Information Processing Systems*, 35:10309–10324.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. (2013). Domain adaptation under target and conditional shift. In *International conference on machine learning*, pages 819–827. PMLR.

Tables and Figures

Table 1: Existing evidence of how PMT performance decreases over time

Paper	Setting	Time Horizon	Decay Estimate	Limitations
Panel A: Model Decay				
Brown et al. (2018)	Ethiopia, Malawi, Nigeria, Tanzania, and Uganda	One or two years	4 percentage points increase in inclusion and exclusion error rates for the poorest 20% of households. 5 percentage points for the poorest 40% of households.	Evaluation metrics are not calculated out-of-sample for the no-decay case, so decay values may be overestimated.
Klasen and Lange (2015)	Bolivia	Three years	No impact on inclusion or exclusion error rates.	Assessment only in a single country using two survey waves.
Panel B: Combined Decay[13]				
Brown et al. (2018)	Ethiopia, Malawi, Nigeria, Tanzania, and Uganda	One or two years	9 percentage points increase in inclusion and exclusion error rates for the poorest 20% of households. 7 percentage points for the poorest 40% of households.	Evaluation metrics are not calculated out-of-sample for the no-decay case, so decay values may be overestimated.
Hillebrecht et al. (2023)	Burkina Faso	One year and three years	2 percentage points increase in inclusion and exclusion error rates for the poorest 30% of households per year.	Assessment only at short time horizons and in a single country using two survey waves; no decomposition into model and data decay.
Sebastian et al. (2018)	Sri Lanka	Three months, six months, and nine months	No evidence of quarter-to-quarter increases in errors of inclusion or exclusion.	Quarter-to-quarter decay is of limited relevance to decay at the time horizon of several years; no decomposition into model and data decay.

Notes: Summary of results from four previously published papers that study data decay, model decay, or combined decay in the context of PMTs. Methadological limitations and conflicting results highlight the need for a systematic and cross-country study of the impacts of model and data recency on PMT efficacy.

[13]To our knowledge no existing papers study data decay alone; only in conjunction with model decay.

Table 2: Panel surveys used in our analysis

Country	**Survey Waves**	**Households**	**Social Registry Covariates**
Ethiopia	**Three** waves over **five** years: 2011, 2013, 2015	3,169	52
Ghana	**Three** waves over **eight** years: 2009, 2013, 2017	3,393	54
Nigeria	**Four** waves over **nine** years: 2010, 2012, 2015, 2018	1,237	48
Peru	**Five** waves over **five** years: 2015, 2016, 2017, 2018, 2019	1,575	28
Tanzania	**Five** waves over **eleven** years: 2008, 2010, 2012, 2014, 2019	424	68
Uganda	**Five** waves over **seven** years: 2009, 2010, 2011, 2013, 2015	1,041	32

Notes: Summary statistics on the six panel surveys used throughout our analysis.

Table 3: Estimates of combined decay

	Baseline Performance	**Combined Decay**
R^2	0.5053	-0.0610***
		(0.0064)
Spearman	0.6923	-0.0287***
		(0.0040)
TER(10)	0.6134	0.0173***
		(0.0021)
TER(20)	0.4667	0.0186***
		(0.0021)
TER(30)	0.3698	0.0170***
		(0.0017)
TER(40)	0.2947	0.0148***
		(0.0016)

Notes: Average total yearly decay when a PMT is allowed to go out of date (neither the data nor the model are updated). Baseline performance represents average PMT performance when there is no lag between data collection, model updating, and PMT evaluation. Decay is calculated using a linear regression of accuracy loss on years elapsed since model and data updating (see Equation 1). * indicates $p < 0.01$, ** indicates $p < 0.05$, and *** indicates $p < 0.01$.

Table 4: Decomposition of combined decay into model decay and data decay

	Baseline Performance	**Data Decay Component**	**Model Decay Component**	**Interaction Term**
R^2	0.5053	-0.0544***	-0.0479***	0.0058***
		(0.0103)	(0.0103)	(0.0021)
Spearman	0.6923	-0.0347***	-0.0115***	0.0033***
		(0.0033)	(0.0033)	(0.0007)
TER(10)	0.6134	0.0200***	0.0080***	-0.0021***
		(0.0018)	(0.0018)	(0.0004)
TER(20)	0.4667	0.0207***	0.0078***	-0.0020***
		(0.0018)	(0.0018)	(0.0004)
TER(30)	0.3698	0.0185***	0.0065***	-0.0015***
		(0.0014)	(0.0014)	(0.0003)
TER(40)	0.2947	0.0167***	0.0057***	-0.0014***
		(0.0013)	(0.0013)	(0.0003)

Notes: Decomposing decay into model decay (decline in acccuracy due to lack of ML model recalibration) and data decay (decline in accuracy due to out-of-date social registry covariates). Baseline performance represents average PMT performance when there is no lag between data collection, model updating, and PMT evaluation. Decay is calculated using a linear regression of accuracy loss on years elapsed since model updating, years elapsed since data updating, and the interaction of the two lags (see Equation 2). We leave data points from Ethiopia out of the R^2 specifications as outliers in Ethiopia's household data cause catestrophically low R^2 values in some cases that bias the overall regression line. Stars are determined by statistical significance of the coefficient in the relevant regression specification: * indicates $p < 0.10$, ** indicates $p < 0.05$, and *** indicates $p < 0.01$.

Table 5: Recommended updating policies for selected real-world social protection systems

Country	Programs	Budget (2015 PPP)	Registry Coverage	Estimated HH in Registry	Budget per Registry HH (2015 PPP)	Recommended Years Between Updates
Argentina	Asignacion Universal por Hijo para Protección Social	$3,743,541,321	99.2%	12,743,127	**$261**	2
Benin	Projet de Services Décentralisés Conduits par les Communauté	$5,963,783	12.3%	260,760	**$20**	10
Colombia	Mas Familias en Accion	$2,065,901,347	72.5%	9,989,794	**$184**	3
	Red Unidos					
Congo	Lisungi	$47,864,073	3.3%	36,614	**$1,163**	1
Costa Rica	Avancemos, Régimen No Contributivo de Pensiones	$493,965,001	12.1%	165,806	**$2,651**	1
	Régimen No Contributivo de Pensiones					
Ecuador	Bono de Desarrollo Humano	$2,419,336,463	47.3%	2,030,637	**$1,060**	1
	Desnutricion Cero					
	Pension para Adultos Mayores					
El Salvador	Comunidades Solidarias Rurales	$127,036,688	10.0%	156,722	**$721**	1
Ghana	Livelihood Empowerment Against Poverty programme	$48,242,902	5.1%	405,211	**$106**	4
Jamaica	Programme of Advancement through Health and Education	$83,760,340	10.3%	89,274	**$835**	1
Mexico	Prospera, Programa de Inclusion Social	$8,924,437,508	15.8%	5,423,243	**$1,465**	1
Peru	Juntos, Pension 65	$1,226,796,360	85.3%	6,999,474	**$156**	3
Philippines	Pantawid Pamilyang Pilipino Program	$3,617,522,838	75.6%	16,569,851	**$194**	3
Togo	Projet de Développement des Communautés et de Filets de Sécurité	$28,238,491	23.4%	365,635	**$69**	6
Median		**$493,965,001**	**15.8%**	**4,289,474**	**$261**	2

Notes: Social protection program data are taken from the Manchester Social Assistance Database (Barrientos, 2018). Data on social registry coverage levels are taken from Grosh et al. (2022), Berner and Van Hemelryck (2021), Beegle et al. (2018), and Leite et al. (2017). Data are presented for all PMT-targeted programs in the year 2015 in the Manchester Social Assistance Database (the most recent year available in the database) for which social registry coverage data are available in one of our sources. Where countries run multiple PMT-targeted programs in the database, the budgets for all such programs are summed together. Where social registry coverage data are available from multiple sources, the most recent source is used. We assume 11% of each program's budget goes towards administrative costs (Ortiz et al., 2017). The estimated number of households in each registry are calculated based on population data from the World Bank (World Bank, 2023), average household size information from the United Nations (Nations, 2017), and the information on registry coverage. Costs are converted to local currency (with exchange rates from the World Bank (World Bank, 2023)), to PPP, and then deflated to 2015 PPP based on the US GDP deflator (PPP exchange rates and GDP deflator are taken from the IMF Economic Outlook Database (IMF, 2023)). The recommended updating frequencies are based on our policy choice simulations (Figure 2), for a coverage level of 30% (although recommended strategies are generally fairly robust across coverage levels). For all real world programs the recommended updating policy is symmetrical (that is, the ML model is recommended to be recalibrated at the same frequency as PMT sweeps), though non-symmetrical updating policies may be recommended for alternative benefit levels (see Figure 2).

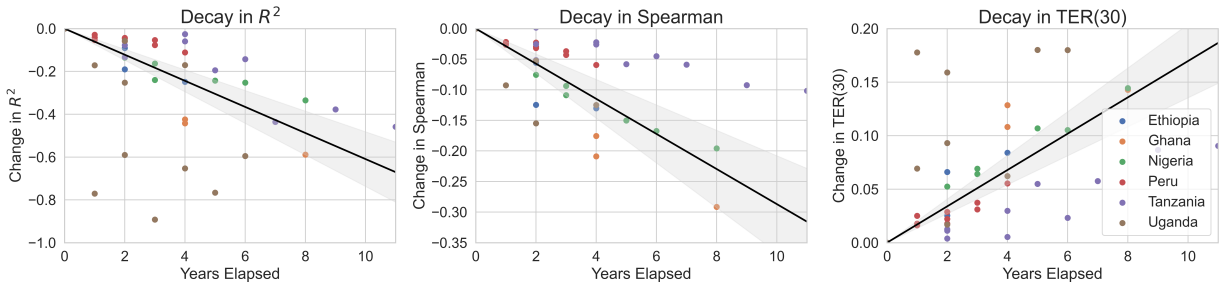

Figure 1: Linear combined decay estimates (Equation 1) are given for the coefficient of determination (left), the Spearman rank correlation coefficient (middle), and the targeting error rate with 30% coverage (right). Individual data points — that is, performance metric differences resulting from specific time gaps between particular training and evaluation rounds — are plotted as points, colored by country. The decay curve from Table 3 is plotted in black, with the 95% confidence interval shown in light grey.

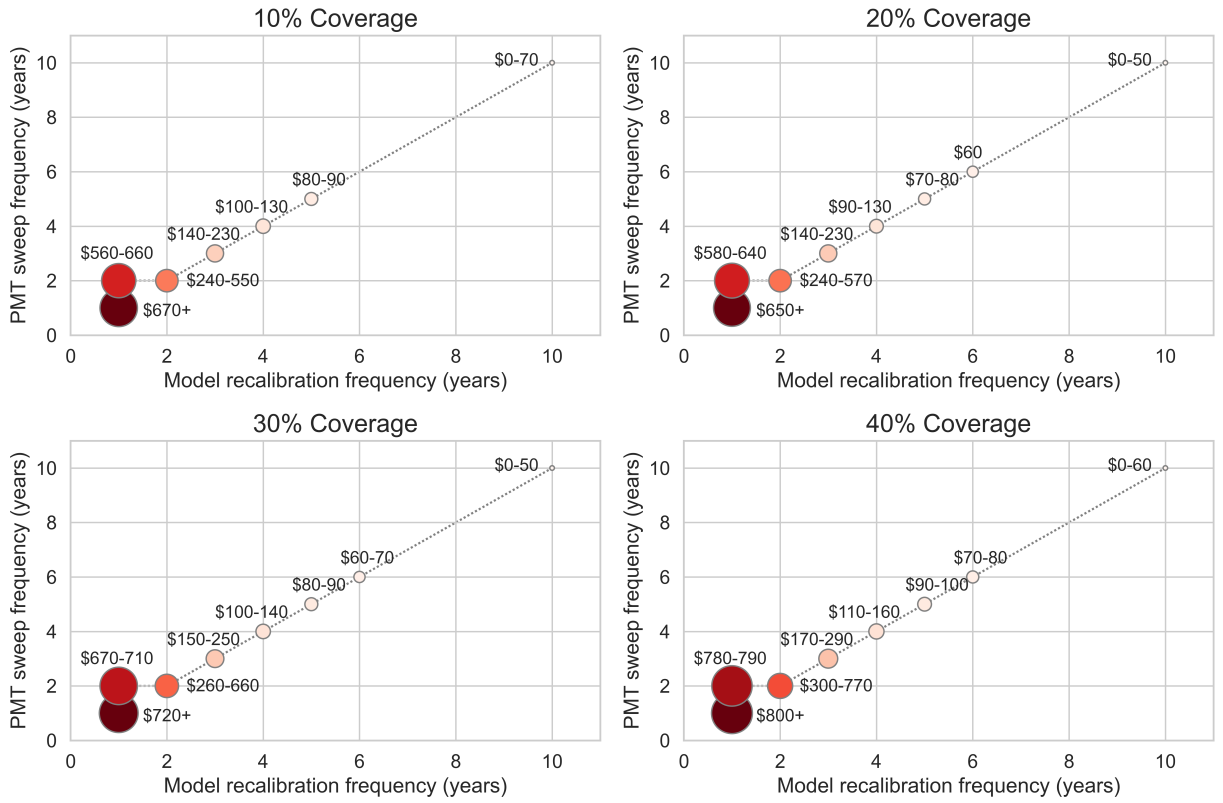

Figure 2: Summary of optimal PMT updating strategies for different program coverage levels (each shown in a different subplot) and different average values of benefits per household in the social registry (shown in colored markers, where markers are colored and sized by the average benefit value). The X-Y location of the marker on each plot denotes the optimal policy, with the model recalibration frequency (in years) on the x-axis and the frequency of PMT sweeps to collect social registry data (also in years) on the y-axis.

A Supplementary tables

Table S1: Estimates of PMT survey costs

Country	Year	Cost per Survey (USD Nominal)	Cost per Survey (2015 PPP)	Source
Burkina Faso	2016	$5.69	$15.52	Schnitzer and Stoeffler (2022)
Chad	2016	$9.50	$22.58	Schnitzer and Stoeffler (2022)
Honduras	2008	$2.62	$6.46	Karlan and Thuysbaert (2019)
Indonesia	2009	$2.70	$9.93	Alatas et al. (2012)
Mali	2016	$4.00	$11.11	Schnitzer and Stoeffler (2022)
Niger	2016	$6.80	$15.52	Schnitzer and Stoeffler (2022)
Peru	2010	$3.05	$6.32	Karlan and Thuysbaert (2019)
Tanzania	2017	$12.00	$29.38	Rosas et al. (2016)
Median		**$4.85**	**$13.31**	

Notes: Cost per household in the social registry for PMT surveys, from four sources. Survey costs are converted to local currency (with exchange rates from the World Bank Development Indicators (World Bank, 2023)), to PPP, and then deflated to 2015 PPP based on the US GDP deflator (PPP exchange rates and GDP deflator are taken from the IMF Economic Outlook Database (IMF, 2023)).

Table S2: Estimates of consumption survey costs

	Total cost (USD Nominal)	**Year**	**Number of HH in Country**	**Cost per HH in Registry (USD Nominal)**	**Cost per per HH in Registry (2015 PPP)**
Afghanistan	$2,289,000	2014	4,125,000	$2.64	$8.97
Bangladesh	$793,600	2010	32,975,809	$0.11	$0.40
Colombia	$1,936,000	2014	13,336,556	$0.69	$1.20
Costa Rica	$1,436,391	2006-2012	1,353,312	$5.05	$8.67
Ethiopia	$1,313,739	2011	20,532,887	$0.30	$1.13
Guatemala	$1,559,790	2014	3,188,816	$2.33	$4.83
Iraq	$3,874,000	2012	4,397,980	$4.19	$9.76
Kyrgyzstan	$245,784	2003	1,200,786	$0.97	$7.59
Malawi	$2,441,929	2010	3,365,799	$3.45	$8.09
Myanmar	$295,200	2015	12,258,083	$0.11	$0.42
Nepal	$1,233,528	2010	6,173,083	$0.95	$3.29
Nicaragua	$773,906	2014	1,193,976	$3.09	$7.97
Niger	$1,188,000	2011	3,757,198	$1.51	$3.36
Nigeria	$1,995,896	2010	35,970,814	$0.26	$0.59
Tanzania	$1,008,885	2014	10,370,317	$0.46	$1.02
Uganda	$1,178,100	2008	6,683,515	$0.84	$2.62
Peru	$2,275,216	2009	7,691,993	$1.41	$3.30
Yemen	$4,291,200	2014	4,142,284	$4.93	$10.87
Median	**$1,375,065**		**5,285,532**	**$1.19**	**$3.33**

Notes: Details of consumption survey cost calculations. Data on survey costs are based on data from the LSMS (Kilic et al., 2017). Data on the number of households in each country are based on population size are from the World Bank (World Bank, 2023), and average household size information from the United Nations (Nations, 2017). The cost per household in a hypothetical social registry is based on a median global social registry coverage of 21% from Grosh et al. (2022) Costs are converted to local currency (with exchange rates from the World Bank (World Bank, 2023)), to PPP, and then deflated to 2015 PPP based on the US GDP deflator (PPP exchange rates and GDP deflator are taken from the IMF Economic Outlook Database (IMF, 2023)).

Table S3: Pre-decay performance of ML models

	Stepwise + LR	**LR**	**LASSO**	**Random Forest**	**Gradient Boosting**
R^2	0.4810	0.4535	0.5275	0.5237	0.5523
Spearman	0.6923	0.6952	0.7193	0.7136	0.7314
TER(10)	0.6134	0.6026	0.5946	0.5831	0.5775
TER(20)	0.4667	0.4592	0.4467	0.4516	0.4395
TER(30)	0.3698	0.3661	0.3526	0.3577	0.3479
TER(40)	0.2947	0.2933	0.2792	0.2833	0.2753

Notes: Average performance for each machine learning model without model or data decay. Average performance across all survey waves is shown.

Table S4: Estimates of combined decay for different ML models

	Stepwise + LR	**LR**	**LASSO**	**Random Forest**	**Gradient Boosting**
R^2	-0.0610***	-0.0682***	-0.0580***	-0.0597***	-0.0645***
	(0.0064)	(0.0068)	(0.0066)	(0.0059)	(0.0064)
Spearman	-0.0287***	-0.0307***	-0.0293***	-0.0308***	-0.0324***
	(0.0040)	(0.0039)	(0.0041)	(0.0042)	(0.0043)
TER(10)	0.0173***	0.0180***	0.0166***	0.0162***	0.0198***
	(0.0021)	(0.0023)	(0.0022)	(0.0027)	(0.0027)
TER(20)	0.0186***	0.0200***	0.0184***	0.0204***	0.0229***
	(0.0021)	(0.0023)	(0.0023)	(0.0022)	(0.0023)
TER(30)	0.0170***	0.0185***	0.0181***	0.0187***	0.0193***
	(0.0017)	(0.0019)	(0.0019)	(0.0020)	(0.0020)
TER(40)	0.0148***	0.0165***	0.0155***	0.0151***	0.0163***
	(0.0016)	(0.0016)	(0.0017)	(0.0018)	(0.0018)

Notes: Replication of results on combined decay in Table 3 for five different ML approaches. Stars are determined by statistical significance of the coefficient in the relevant regression specification: * indicates $p < 0.10$, ** indicates $p < 0.05$, and *** indicates $p < 0.01$.

Table S5: Estimates of combined decay for each country separately

	Pre-decay performance						**Combined decay**					
	Ethiopia	Ghana	Nigeria	Peru	Tanzania	Uganda	Ethiopia	Ghana	Nigeria	Peru	Tanzania	Uganda
R^2	0.3026	0.5600	0.4983	0.7119	0.3469	0.4299	-0.0649***	-0.0852***	-0.0468***	-0.0259***	-0.0395***	-0.1389***
							(0.0066)	(0.0073)	(0.0033)	(0.0017)	(0.0039)	(0.0247)
Spearman	0.5537	0.7646	0.7176	0.8473	0.6270	0.6220	-0.0368***	-0.0403***	-0.0278***	-0.0149***	-0.0092***	-0.0844***
							(0.0052)	(0.0027)	(0.0012)	(0.0008)	(0.0007)	(0.0133)
TER(10)	0.6655	0.5159	0.6372	0.4586	0.7615	0.6282	0.0299***	0.0275***	0.0232***	0.0184***	0.0039**	0.0396***
							(0.0023)	(0.0027)	(0.0012)	(0.0019)	(0.0015)	(0.0069)
TER(20)	0.5255	0.3863	0.4554	0.3017	0.5949	0.5257	0.0334***	0.0246***	0.0241***	0.0178***	0.0063***	0.0419***
							(0.0033)	(0.0032)	(0.0009)	(0.0022)	(0.0008)	(0.0075)
TER(30)	0.4444	0.3118	0.3475	0.2286	0.4517	0.4368	0.0216***	0.0218***	0.0192***	0.0128***	0.0078***	0.0390***
							(0.0027)	(0.0026)	(0.0007)	(0.0007)	(0.0006)	(0.0059)
TER(40)	0.3663	0.2481	0.2695	0.1938	0.3378	0.3580	0.0136***	0.0194***	0.0160***	0.0096***	0.0064***	0.0376***
							(0.0025)	(0.0018)	(0.0005)	(0.0011)	(0.0006)	(0.0049)
N	3	6	3	6	4	10	5	15	5	15	5	15

Notes: Pre-decay performance and estimate of combined decay calculated for each country separately.

B Assessing model and data refresh policies based on social welfare

An alternative approach to calculate the optimal PMT refresh policy (that is, alternative to the approach we take in the main paper) is to reason about social welfare: assume that a social protection program has a certain budget, and that any of that budget that goes towards survey costs is budget that will not go towards benefits delivered to beneficiaries. In this framework, the welfare impacts of a program are determined by the size of the benefits and who the benefits are targeted to — thus more up-to-date PMTs yield higher social welfare through better targeting, but lower social welfare through lower benefit amounts. The best PMT refresh policy in this framework is the the one that yields that greatest improvements in total social welfare.

B.1 Calculating total social welfare

We start by calculating the total social welfare for each targeting scenario we simulate. Specifically, we assume a fixed benefit size \$$b$, that goes to each eligible household (measured in 2015 USD PPP). We also fix a program coverage rate k% (from {10%, 20%, 30%, 40%}). For each lagged train-test pair in our dataset, we simulate providing the PMT-identified poorest k% of households in the test set with benefits \$$b$, and calculate the aggregate utility post-benefits ($U_{program}$) using the CRRA utility function (Hanna and Olken, 2018).[14] To make the units of utility interpretable and comparable across validation sets of different sizes, we transform aggregate utility to a measure of improvement over the status quo, by calculating the percent change in the status-quo (pre-benefits) utility (U_{before}) as a result of the targeted aid program:

$$\text{Utility improvement} = \frac{U_{program} - U_{before}}{U_{before}} \tag{4}$$

Naturally, the utility impact depends critically not just on the delay since model recalibration and social registry data collection, but also on the size of the benefits delivered to each household and the coverage of the program. For a sense of the magnitude of utility impacts at different benefit sizes and coverage levels, Table S6 records average utility impacts for sixteen hypothetical social protection programs spanning the four coverage levels and four different benefit sizes, in the no-decay setting.

[14] Following Hanna and Olken (2018), we use $\rho = 3$ in the CRRA utility function.

B.2 Parameterizing utility decay

Next, we parameterize social welfare decay as a function of time since model recalibration and time since covariate data collection, just as we do for other metrics in the main paper (but with the addition of an intercept, since here our outcome variable is total utility impact, not the change in utility impact from the no-decay setting). For each the four program coverage settings tested and a wide grid of possible benefit sizes, we regress the utility improvement measure above on the time since model calibration, the time since covariate data collection, and the interaction of the two terms. Figure S1 plots combined utility decay, and Table S7 decomposes combed utility decay into model decay and data decay.

B.3 Finding the best updating policy

Finally, we simulate the same 100 updating policy combinations as previously of recalibrating the model every 1-10 years and conducting a PMT sweep every 1-10 years. For each option we calculate the total yearly survey cost, which consists of the sum of average yearly consumption survey costs and average yearly PMT sweep costs. Of the total budget, 11% is assumed to go to administrative costs (Ortiz et al., 2017), and after accounting for survey costs, the remaining yearly budget is assumed to go to benefits: so, for example, a program that has a budget of $50 PPP per household in the registry and updates both the model and social registry data each year will have provide approximately $25 PPP per household in the registry in benefits. After fixing a program coverage level and determining the amount of benefits each policy will provide yearly, we calculate the benefits that will go to targeted households, and then use the parameterization of utility decay above to calculate expected average yearly social welfare over the life cycle of the program. Finally, we identify the updating policy with the highest expected social welfare.

The best updating policies identified through this process are shown in Figure S2. In summary, we find results slightly more in favor of frequent surveys than those in the main paper. Using the social welfare framework, programs with a budget below $40 PPP per households in the social registry should recalibrate the PMT model every year and conduct a PMT sweep every 2-3 years. Programs with a budget above $40 PPP per registry household should recalibrate the PMT model and conduct a PMT sweep every year.

Table S6: Examples of no-decay welfare impacts

	$10 benefits	**$100 benefits**	**$500 benefits**	**$1000 benefits**
10% coverage	0.2%	5.4%	14.3%	18.9%
20% coverage	1.5%	10.0%	24.1%	31.5%
30% coverage	2.3%	13.0%	31.0%	40.5%
40% coverage	2.8%	14.9%	35.9%	47.1%

Notes: Average social welfare impact (improvement over the status quo) for four different coverage levels and four different benefit sizes. Benefit sizes are measured in USD 2015 PPP and represent the yearly benefits to targeted households (not the benefits per household in the registry).

Table S7: Parameterization of utility decay

	Data decay	**Model decay**	**Interaction**	**Intercept**
$100 benefits	-0.0124***	-0.0118***	0.0013*	0.1754***
	(0.0041)	(0.0041)	(0.0008)	(0.0139)
$500 benefits	-0.0236***	-0.0200***	0.0025**	0.3839***
	(0.0063)	(0.0063)	(0.0012)	(0.0212)
$1,000 benefits	-0.0256***	-0.0201***	0.0028**	0.4724***
	(0.0061)	(0.0061)	(0.0012)	(0.0207)

Notes: Parameterization of utility decay for a 30% coverage program at three hypothetical benefit amounts: $100, $500, and $1,000. Benefit amounts are measured in USD 2015 PPP and represent the yearly benefits to targeted households (not the benefits per household in the registry).

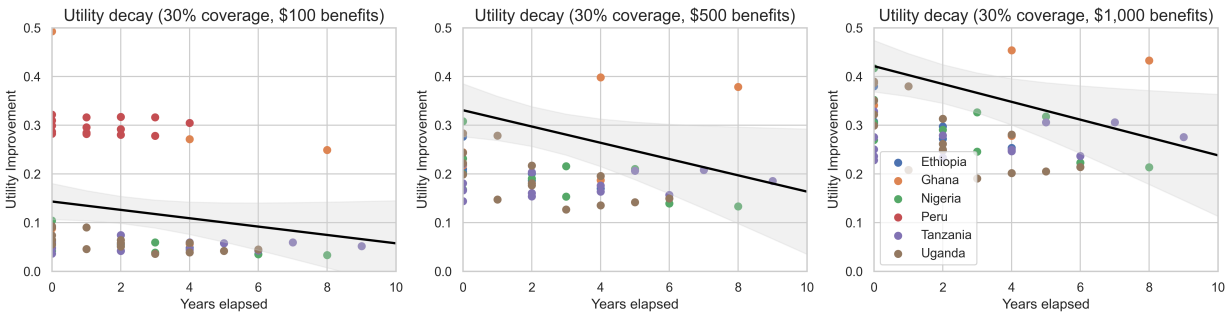

Figure S1: Linear estimates of combined utility decay are given for the coefficient of determination (left), the Spearman rank correlation coefficient (middle), and the targeting error rate with 30% coverage (right). Individual data points — that is, performance metric differences resulting from specific time gaps between particular training and evaluation rounds — are plotted as points, colored by country. The combined decay curve is plotted in black, with the 95% confidence interval shown in light grey.

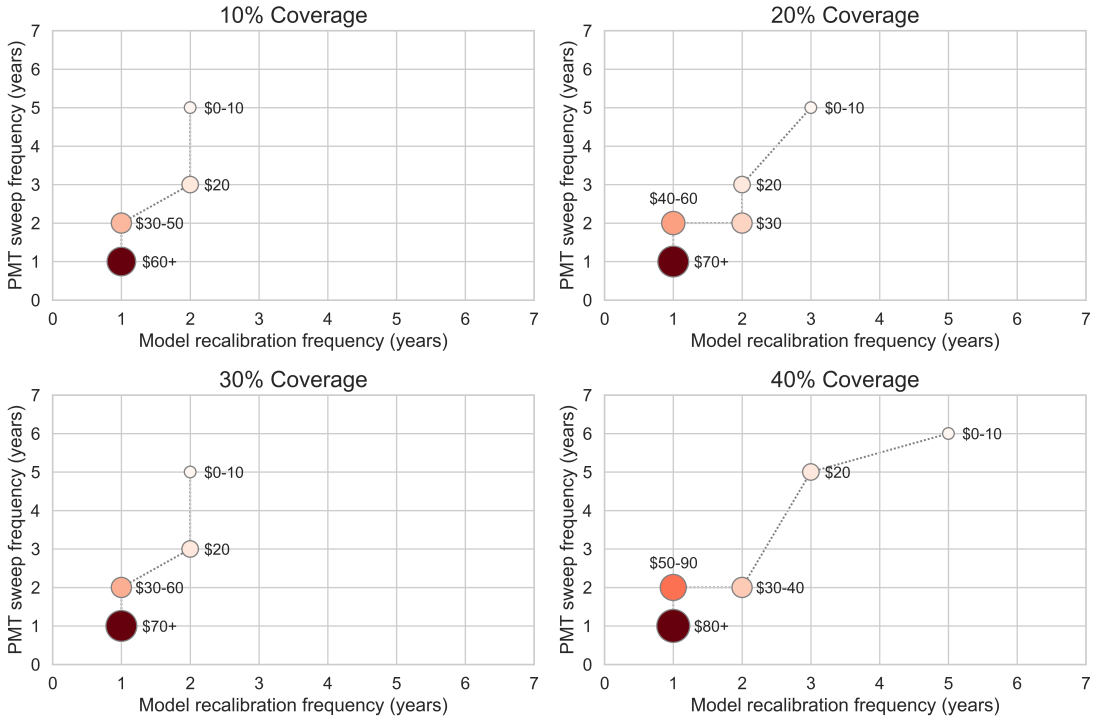

Figure S2: Summary of optimal PMT updating strategies optimized by total utility. Results are shown for different program coverage levels (each shown in a different subplot) and different average values of benefits per household in the social registry (shown in colored markers, where markers are colored and sized by the average benefit value). The X-Y location of the marker on each plot denotes the optimal policy, with the model recalibration frequency (in years) on the x-axis and the frequency of PMT sweeps to collect social registry data (also in years) on the y-axis.