Targeting Impact versus Deprivation

By Johannes Haushofer, Paul Niehaus, Carlos Paramo, Edward Miguel, and Michael Walker*

A large literature has examined how best to target anti-poverty programs to those most deprived in some sense (e.g., consumption). We examine the potential tradeoff between this objective and targeting those most impacted by such programs. We work in the context of an NGO cash transfer program in Kenya, employing recent advances in machine learning methods and dynamic outcome data to learn proxy means tests that jointly target both objectives. Targeting solely on the basis of deprivation is not attractive in this setting under standard social welfare criteria unless the planner's preferences are extremely redistributive.

JEL: O12, I32, D63

Keywords: targeting; cash transfers; machine learning; social welfare

Targeting is a core element of anti-poverty program design in both poor and rich countries, with program benefits typically targeted to those households or individuals who are "deprived" in some sense, for instance, in terms of wealth, income, or living standards. There is a growing literature in development economics focused on how effectively one can identify such deprived households to target them with anti-poverty programming, via proxy means tests (PMT), com-

* Haushofer: Department of Economics, National University of Singapore and Stockholm University, 1 Arts Link AS2 #06-02 Singapore 117570 (e-mail: johannes.haushofer@nus.edu.sg); Niehaus: Department of Economics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093 (e-mail: pniehaus@ucsd.edu); Paramo: Department of Economics, University of California, Berkeley, 530 Evans Hall, Berkeley, CA 94720-3880 (carlos.paramo@berkeley.edu); Miguel: Department of Economics, University of California-Berkeley, 549 Evans Hall #3880, Berkeley, CA 94720-3880, and NBER (e-mail: emiguel@econ.berkeley.edu); Walker: University of California, Berkeley Center for Effective Global Action 207 Giannini Hall Berkeley, CA 94720 (e-mail: mwwalker@berkeley.edu). We thank Siwan Anderson, Arun Chandresekhar, Rajeev Dehejia, Matthew Gentzkow, Larry Katz, David McKenzie, Jonathan Morduch, Suresh Naidu, Ben Roth, Vira Semenova, Monica Singhal, Jack Willis, Kaspar Wüthrich and conference and seminar audiences at the 2021 ASSA and LACEA meetings, BREAD, CEGA, CEPR Political Economy, Columbia, Harvard, National University of Singapore, NYU, Oxford CSAE conference, Stanford, University of Auckland, University of British Columbia, University of Minnesota, UC Merced, UC San Diego, and USC for helpful comments and suggestions. We thank Justin Abraham, Aakash Bhalothia, Christina Brown, Genevieve Deneoux, Tilman Graff, Grady Killeen, Maximiliano Lauletta, Michelle Layvant, Layna Lowe, Anya Marchenko, Gwyneth Miner, Max Mueller, Priscilla de Oliviera, Robert On, Rachel Pizatella-Haswell, Nachiket Shah, Emaan Siddique, Zenan Wang, Francis Wong and Kejian Zhao for excellent research assistance. Niehaus is a co-founder, former president (2012-17) and current board member of GiveDirectly. This research was supported by grants from the National Science Foundation, International Growth Centre, CEPR/Private Enterprise Development in Low-Income Countries (PEDL), the Weiss Family Foundation, and an anonymous donor. The computations and data handling were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at NSC partially funded by the Swedish Research Council through grant agreement no. 2018-05973. The study received IRB approval from Maseno University (MSU/DRPC/MUERC/000090/14) and the Univ. of California, Berkeley (Protocol 2014-05-6354). AEA Trial Registry RCT ID: AEARCTR-0000505, Egger et al. (2024).

munity input, ordeal mechanisms, "big data", and other approaches (see Alatas et al., 2012; Blumenstock, Cadamuro and On, 2015; Brown, Ravallion and van de Walle, 2018; Hanna and Olken, 2018, among others).

Yet we know conceptually that targeting the most deprived is potentially only part of the problem facing a social planner or policymaker. Welfare-maximizing allocations of scarce resources should generally depend both on how poor people are to begin with and also on how much they would benefit from receiving additional assistance. As a mechanical example, targeting small business skills training to people who are unable (for any reason) to themselves run a business would not yield economic gains, and so would be a waste of resources. This basic point is analogous to issues that arise generically in many other policy contexts—whether, for example, to triage limited health care resources to the sickest patients versus to those deemed most likely to recover, or a classroom teacher's time to the worst performing pupils versus toward generating the most value added. Its implication is that we can safely focus solely on targeting deprived households when treatment effect magnitudes are similar for everyone but not when they vary meaningfully.[1]

This is not an idle concern, as there is growing evidence that the effects of some important interventions do vary meaningfully. Heterogeneous treatment effects are empirically important in the recent microcredit literature in development economics, for example (Banerjee, Karlan and Zinman, 2015; Meager, 2022)—perhaps because the "poorest of the poor" sometimes lack the circumstances or complementary inputs and skills to successfully invest their loans. Indeed, these barriers may be part of the reason they are poor to begin with. Similarly, Bhattacharya and Dupas (2012) show that the effects of a subsidy for purchasing insecticide-treated bednets vary predictably, and that one could meaningfully increase average effects on utilization by exploiting this variation. Such findings raise the question to what extent there is an impact/deprivation trade-off in targeting anti-poverty programs.

This tension is likely to be particularly relevant for cash transfers, an increasingly common form of anti-poverty programming (see Baird, McIntosh and Özler, 2011; Haushofer and Shapiro, 2016; Bastagli et al., 2016, among many others). The intrinsic flexibility of cash means that different households can use it in distinct ways, some of which policymakers and planners may prefer to others—targeting a high (or low) marginal propensity to consume, for example. Even household preferences that are homogeneous but non-homothetic will mechanically lead to differential patterns of impact across poor and rich households. And different households may not face the same constraints, including both "internal constraints" in the form of behavioral biases, and external constraints such as credit market failures, which are thought to be pervasive in low- and middle-

[1] Other important political economy considerations regarding program targeting—influencing voting, for instance (see Lindbeck and Weibull, 1987; Manacorda, Miguel and Vigorito, 2011)—are not our focus in this paper.

income countries (LMICs). Factors such as these likely contribute to the substantial heterogeneity actually observed in the impacts of cash transfers (e.g., de Mel, McKenzie and Woodruff, 2008; Hussam, Rigol and Roth, 2022).

This paper characterizes and quantifies the trade-off between targeting deprivation and impact in the context of a large-scale unconditional cash transfer program in rural Kenya. This setting, previously described in Egger et al. (2022*b*) (henceforth, EHMNW), has several characteristics valuable for this purpose. The transfer program targeted a relatively large share (35%–40%) of households in treated villages using a simple PMT, allowing us to consider optimal targeting within an unusually representative share of the population. Data collection covered a relatively large sample of 4,749 transfer-eligible households, allowing us to use data-intensive statistical techniques. And the timing of both treatment onset and outcome measurement was experimentally varied. As we discuss below, this allows us to account for dynamics in the analysis rather than making the implausible assumption that household outcomes stay constant over time, which poverty-targeting analyses have often been forced to adopt (due to lack of data such as ours).

We proceed in two steps. We first predict for each household both its time-averaged treatment effect if treated, and its time-averaged deprivation if not. We do so using a common set of "PMT-like" baseline characteristics as predictors, so that the exercise holds constant the type of information typically available to real-world program designers. We use a machine learning approach to prediction, as pre-specified on the AEA RCT Registry,[2] building on recent advances optimized specifically for the study of heterogeneous treatment effects, including Wager and Athey (2018), Chernozhukov et al. (2018), and especially the generalized random forest (GRF) estimator of Athey, Tibshirani and Wager (2019). This approach leads to considerable variation in both predicted deprivation and predicted impacts in this sample. We then use this joint distribution to identify and characterize the groups that a planner maximizing a canonical social welfare function would choose to target.[3] We emphasize functional forms that smoothly parameterize the strength of the planner's redistributive preferences, while also considering as benchmarks the extreme cases of targeting based solely on deprivation or on impact. We apply this approach to a set of pre-specified financial outcomes—consumption expenditures, assets, and income—that are important objectives for development policymakers, as well as to measures of food security.[4]

The main substantive result is that—across a wide range of social welfare functions and degrees of redistributive preference—the planner consistently takes both

[2] See `https://www.socialscienceregistry.org/trials/505` for more information.

[3] A two-step approach is necessary to examine the policy-relevant trade-off between deprivation and impact that is central to this paper. That said, a one-step analogue, as studied in the theoretical literature on Empirical Welfare Maximization (Manski, 2004; Kitagawa and Tetenov, 2018; Athey and Wager, 2021), might yield benefits in terms of statistical efficiency.

[4] As Sen (1999) articulated, choices of space like these are consequential; results might differ if we focused on equity and efficiency in some other space, e.g. of capabilities.

deprivation and impact into account. That is, they choose to target both a substantial share of households that are not among the most impacted, and a substantial share of those that are not among the most deprived. For our preferred constant absolute risk aversion (CARA) utility function, we reject the null of (approximately) complete overlap with the most deprived group for any plausible curvature value. It is not until we reach extreme levels of curvature—levels that effectively place over 360 times as much weight on a household with half the per capita consumption of another—that targeting based solely on deprivation is approximately optimal. The same broad conclusion continues to hold using other functional forms (e.g. CRRA); with the introduction of aggressive time discounting; in a simulated second round of (re-)targeting; and if we incorporate welfare weights estimated from data on actual community decisions collected by Alatas et al. (2012) that capture preferences for privileging disadvantaged groups (specifically, widows), using either the standard Pareto approach or the linear approach proposed by Saez and Stantcheva (2016). Across all these specifications we strongly reject the nulls that the planner should target based solely on deprivation or based solely on impact.

These conclusions reflect the fact that there is a meaningful trade-off in this sample between targeting for deprivation versus for impact. Taking household consumption as a leading example, those predicted to be in the most deprived half of the sample if untreated do indeed have lower time-averaged per capita consumption (by 45%) than those predicted to be in the most impacted half. However, the time-averaged treatment effect on consumption is 64% larger in the most impacted half of the sample compared to the most deprived half. A similar trade-off holds for other outcomes, though the magnitudes differ somewhat, indicating that the trade-offs facing policymakers may also depend on the key outcome of interest.[5]

From the perspective of theories of poverty dynamics, several characteristics of the deprived and impacted groups are noteworthy. Household size is the most important predictor, by a wide margin, of both deprivation and impact, with larger households both benefiting more from treatment—echoing classic ideas of scale economies in household production (Nelson, 1988; Deaton and Paxson, 1998)—and faring better without it. (The net result is that larger households are more likely to be targeted by the planner.) This appears to be closely linked to lifecycle patterns, as households headed by middle-aged adults are more likely to be among the most impacted, while those headed by the young and the old are more likely to be deprived. Large treatment effects on financial outcomes do not come

[5] For the pre-specified food security index we do find some evidence that more deprived households experience larger treatment effects, suggestive of a "hierarchy of needs" (as in Maslow, 1943). But this index—akin to those commonly used in development economics and based on survey responses regarding lack of food—appears to capture per capita rather than total household food consumption. This is problematic since all households received the same amount of money, so that per-capita effects will mechanically tend to be smaller in larger households. As shown below, if we simply examine total consumption of food instead, the patterns again indicate a trade-off between targeting for deprivation versus impact, consistent with trends for the financial outcomes.

at the expense of leisure, or of side transfers or loans to other households, but are positively associated with having applied for a loan in the last year, suggesting a role for credit constraints. Finally, the same households tend to experience larger gains in all three of consumption, assets, and income, and correlations across these three effects are quite stable over time (since receipt of the cash transfer). Economically this suggests that heterogeneity in impacts may reflect differences in investment and market opportunities, broadly construed, more than differences in the propensity to save or invest per se.

One potential caveat to these results involves the role of spillover effects. EHMNW document a sizable transfer multiplier of 2.5 due to the cash transfer program in the study area. The mere existence of spillovers does not necessarily alter the interpretation of the main results: the conclusions would be the same if all households caused and experienced the same additive spillovers, for example (at least for CARA social welfare functions). But the interpretation would change to the extent the results capture predictable differences in which households experience larger spillover effects. In two auxiliary tests for this, using data on both cash transfer eligible and ineligible households and both within- and between-village exposure to treatment, the analysis does not detect meaningful heterogeneity in the impact of spillovers. This may give a greater degree of confidence in the main results.

As a final methodological point of interest, we contrast results obtained using GRF to those obtained using a simple OLS regression—which, while not regularized, has been widely used in the past for prediction in targeting analyses—as well as LASSO, a classic and ubiquitous ML model (Tibshirani, 1996). Neither tool is explicitly designed for learning heterogeneous effects, but both can be adapted to do so. In our data both LASSO and OLS (but especially OLS) select groups that are somewhat less deprived than GRF, and also substantially less impacted, primarily because they mistakenly predict that some households are far more impacted than they truly are. This suggests that it may be prudent to learn heterogeneous effects for targeting applications using recent methods such as GRF designed specifically for that purpose.[6]

A central finding is that the most deprived households should not always and necessarily be the sole focus of anti-poverty program targeting, although that is the norm in practice. The results indicate that there are important trade-offs for policymakers to consider. In this sense they echo, on a microeconomic scale, longstanding debates regarding potential trade-offs between equity and efficiency in the process of economic growth and development more generally (Alesina and Rodrik, 1994; Persson and Tabellini, 1994; Banerjee, Gertler and Ghatak, 2002). They also parallel recent work by Björkegren, Blumenstock and Knight (2022) studying the dual of the problem we study here, i.e. which policymaker prefer-

[6] As we discuss below, the issue here appears to be analogous to that identified by Abadie, Chingos and West (2018), who document a bias in conventional approaches to studying impact heterogeneity towards negative estimates of the relationship between impact and untreated outcomes. In contrast, the GRF approach used here yields positive estimates.

ences rationalize a given observed targeting rule; they infer preferences that value targeting both deprivation and impact.[7,8]

That said, the findings in this study apply to one intervention in a single setting, and one program in isolation. Considering a portfolio of anti-poverty interventions, targeting one towards the most impacted may strengthen the case for targeting others towards the most deprived. For example, an optimal strategy might involve targeting cash transfers to those who benefit most from them (in terms of future income gains), while simultaneously working to remove for the most deprived the barriers that limit their ability to benefit from assistance. Doing so may be particularly important for socially marginalized groups (e.g., female headed households, migrants and members of ethnic or religious minorities) who may lack the same market opportunities as other households.

I. Conceptual framework

We study the problem of choosing which households h to receive treatment (e.g., program assistance) in order to maximize a social welfare function

$$\sum_h \sum_{t=0}^{\bar{t}} \delta^t W(Y_{h,t}(T_h)). \tag{1}$$

Here $Y_{h,t}$ is a real-valued outcome of interest such as consumption, wealth, or food security, which potentially depends on the household's assignment to receive treatment, indicated by $T_h \in \{0,1\}$. For simplicity we will think for now of each household as having a single member, abstracting from variation in household size (which is introduced when we map the framework to the data in Section III). As is standard, the function $W : \mathbb{R} \to \mathbb{R}$ satisfies $W' > 0$ so that higher values of each household's outcome are preferred, and $W'' \leq 0$ so that gains matter (weakly) more for households that are more deprived to begin with. We frame the problem as inherently dynamic: the planner chooses an allocation once but also cares about well-being at other future times. The optimal allocation will generically depend on the full time paths of both $Y_{h,t}(1)$ and $Y_{h,t}(0)$, not just on untreated outcomes $Y_{h,0}(0)$ at the start of the program (as in PMT analyses

[7] Another related, emerging literature in development economics examines the potential trade-off between deprivation and impact across alternative targeting paradigms. Premand and Schnitzer (2021) compare PMT targeting to alternatives in a cash transfer program in Niger and do not find evidence of a trade-off. Basurto, Dupas and Robinson (2020) show that chiefs in Malawi tasked with assisting the needy tend to target productive farm inputs to households that have higher returns to their use, relative to the allocation achieved by a strict PMT approach.

[8] Other recent work examining heterogenous treatment effects of anti-poverty programs using ML methods includes McKenzie and Sansone (2019), who find limited additional benefits from using machine learning methods over and above the predictive power of a few key covariates in predicting entrepreneurial success in Nigeria; Hussam, Rigol and Roth (2022), who examine treatment effects forecasts obtained via machine learning as a benchmark for those elicited from community members; and Bertrand et al. (2021), who employ ML and other approaches to evaluate how to improve the targeting of workfare programs in Ivory Coast.

based on a single baseline survey). We set a finite time horizon $\bar{t}$ to reflect the fact that programs are typically re-targeted every few years. Given this, and for expositional simplicity, we normalize the time discount factor $\delta = 1$ for the rest of this section, but will examine the impacts of discounting in the empirical analysis.

Using potential outcomes notation allows this objective to be rewritten as

$$\sum_h \sum_{t=0}^{\bar{t}} W(Y_{h,t}(T_h)) = \sum_h \sum_{t=0}^{\bar{t}} W(Y_{h,t}^0 + T_h \cdot \Delta_{h,t})) \tag{2}$$

where $Y_{h,t}^{T_h} \equiv Y_{h,t}(T_h)$ and $\Delta_{h,t} \equiv Y_{h,t}^1 - Y_{h,t}^0$ is h's treatment effect. This reformulation highlights the potential tension between two distinct objectives: targeting benefits to those *worst-off* absent the intervention (i.e. have the smallest $Y_{h,t}^0$'s), and targeting benefits to those who will be *most positively impacted* by the intervention (largest $\Delta_{h,t}$'s). These objectives are captured in a disciplined way here, in the sense that both are tightly linked through the function $W(\cdot)$; $W(\cdot)$ determines both the strength of preference for targeting deprived households, and also the extent to which large treatment effects are discounted due to diminishing marginal benefits. One can of course also readily extend the framework by incorporating ad hoc weights to capture other forms of distributive preference; in this case the objective would be

$$\sum_h \sum_{t=0}^{\bar{t}} W(Y_{h,t}(T_h)) = \sum_h \sum_{t=0}^{\bar{t}} \omega_h W(Y_{h,t}^0 + T_h \cdot \Delta_{h,t})) \tag{3}$$

The weights $\{\omega_h\}$ here might reflect a desire to correct structural inequities facing particular subgroups (i.e., by age, gender, or ethnic background) that are not fully captured by low values of Y^0, for example. We abstract from this idea for now, but will reintroduce and illustrate it in the empirical application (see in particular Section IV.B).

The criterion function (2), and in particular the variation in treatment effects, can be interpreted in two distinct ways. One is that $W(\cdot)$ correctly represents households' preferences over their own outcomes, but that households face different opportunities and constraints. Some may possess investment opportunities that others lack, for example, so that they are able to increase their standard of living more after receiving treatment (a household cash transfer in the empirical application). In this case households might agree – from a vantage point behind a "veil of ignorance" in which they do not yet know their specific draw of $(Y_{h,t}, \Delta_{h,t})$ – that (2) is the appropriate objective of policy. Alternatively, $W(\cdot)$ may represent the preferences of a paternalistic planner or policymaker, which differ from those of the households themselves. For example, households' time preferences may vary, and the policymaker may prefer that they make relatively "patient" choices.[9] In this case, maximization of (2) would implement policymaker rather than household preferences.

We consider how to balance the objectives captured by (2) subject to information constraints facing a typical policymaker. Specifically, we suppose that she cannot observe $Y^0_{h,t}$ and $\Delta_{h,t}$ in the full population. This reflects the costs of gathering data on complex outcomes such as consumption, the fact that claims about these outcomes are hard to verify, and (in the case of $\Delta_{h,t}$) the more fundamental issue that she can never directly observe a household's counterfactual outcomes. Instead we suppose that she observes a set of baseline covariates $X_h \in \mathbf{X}$ in the full population, as well as the realized outcomes $Y_{h,t}(T_h)$ from an experimental sub-sample that is representative (possibly after re-weighting) with respect to both households h and time $t \in [0, \bar{t}]$. We think of X_h as representing the kinds of variables typically seen in proxy means tests used to target programs in LMICs, e.g. major assets, household size, number of children, sector of employment, etc. The planner uses these data to select a rule $\mathcal{R} : \mathbf{X} \to \{0, 1\}$ determining assignment to treatment in the rest of the population, subject to any budget or enrollment constraints, for instance, that there is sufficient funding to treat a share ϕ of households in the population.

Data from this experimental sample enable the planner to consider targeting based on predictions

$$\hat{Y}^0_t(X_h) \text{ of } \mathbb{E}[Y^0_{h,t}|X_h, t] \tag{4}$$

$$\hat{\Delta}_t(X_h) \text{ of } \mathbb{E}[Y^1_{h,t} - Y^0_{h,t}|X_h, t] \tag{5}$$

obtained from these data. For a given $W(\cdot)$ the natural approach is to rank households by the incremental contributions to social welfare that treating them would induce given these predictions:

$$d\hat{W}(X_h) \equiv \sum_{t=0}^{\bar{t}} \Big[W(\hat{Y}^0_t(X_h) + \hat{\Delta}_t(X_h)) - W(\hat{Y}^0_t(X_h)) \Big] \tag{6}$$

$$\mathcal{R}^*(X_h) \equiv 1(d\hat{W}(X_h) \geq q^{d\hat{W}}_{1-\phi}) \tag{7}$$

where q^Z_ϕ denotes the ϕ'th percentile of the empirical distribution of given variable Z. This rule $\mathcal{R}^*$ strikes a balance between targeting deprivation and impact, with the terms of the tradeoff governed by the curvature of W. Dynamics may be important if the joint distribution changes over time: for instance, if the most deprived consume most of the transfer initially, while the less-deprived invest more and consume more later, an issue that we explore in the empirical application.[10,11]

[9] Paternalism over others' time preferences seems to be common, as for example Ambuehl, Bernheim and Ockenfels (2021) document in the lab.

[10] That said, the same tradeoffs emerge even if the outcomes the planner cares about are realized only once.

[11] In contrast, the Empirical Welfare Maximization literature (Manski, 2004; Kitagawa and Tetenov,

In contrast, the approach typically used in practice to learn targeting rules is to base them solely on predictions of deprivation at time 0, using treatment rules of the form

$$\mathcal{R}^D(X_h) = 1(\hat{Y}_0^0(X_h) \leq q_\phi^{\hat{Y}}) \tag{8}$$

A known limitation is that deprivation is not stable over time, so that it would be preferable if possible to target based on each household's predicted time-averaged deprivation, i.e. $\hat{Y}^0(X_h) \equiv \sum_t \hat{Y}_t^0(X_h)$, an approach we will implement below. The deeper issue, however, is that targeting based solely on deprivation may miss the opportunity to target high-impact recipients, if treatment effects are not homogenous.

The opposite extreme would be to target based solely on predicted impact:

$$\mathcal{R}^I(X_h) = 1(\hat{\Delta}(X_h) \geq q_{1-\phi}^{\hat{\Delta}}) \tag{9}$$

where predicted time-averaged impact is denoted by $\hat{\Delta}(X_h) \equiv \sum_t \hat{\Delta}_t(X_h)$. This policy is uncommon in practice, to our knowledge. Program evaluation studies increasingly examine it, though these are usually limited to examining treatment effects at a single point in time post-intervention. Generally speaking, targeting impact alone will tend to be appealing if Y_h^0 does not vary (much) relative to Δ_h or if $W(\cdot)$ is (nearly) linear in its argument.

Below we will explore the trade-off between targeting deprivation and impact quantitatively by examining the joint distribution of $(\hat{Y}^0(X_h), \hat{\Delta}(X_h))$ and how the particular households h selected for treatment vary depending on $W(\cdot)$. As benchmarks, and given their prominence in the existing literature, we will also compare the households selected in this way to those selected by the simpler rules that target only deprivation and only impact.

Note that the approach here, where targeting is explicitly grounded in social welfare maximization, allows for more nuance than the more ad hoc approach common in the literature that parametrizes a two-cost loss function, one for errors of inclusion and one for errors of exclusion. An obvious benefit is that the approach here captures, for example, the idea that excluding a household that is only slightly below the poverty line is not as costly as excluding one that is extremely poor. That said, the fact that existing work often places more weight on errors of exclusion (i.e., failing to target a poor household) than on errors of inclusion itself suggests that these studies, too, have an implicit social welfare function in mind.

2018; Athey and Wager, 2021) focuses on predicting $W(Y(T, X))$ using X directly, yielding predictions $\hat{W}_h(T_h)$, and then selecting for treatment observations with high values of $\hat{W}_h(1) - \hat{W}_h(0)$. This approach yields useful guarantees about the asymptotic performance of the targeting rule, but obscures the policy relevant tradeoff between impact and deprivation that we wish to draw out here.

II. Study design

We study targeting in the context of a large-scale experimental evaluation of unconditional cash transfers to low-income rural Kenyan households (Egger et al., 2024), previously examined by EHMNW. That paper provides details on the setting and design which we briefly summarize here.

A. *Setting: rural western Kenya*

The project took place in three contiguous subcounties of Siaya County, a largely rural area in western Kenya, which the NGO GiveDirectly (GD) had selected based on its high poverty levels. Within this area, GD selected rural (i.e., not peri-urban) villages in which it had not previously worked. This yielded a final sample of 653 villages spread across 84 sublocations (the administrative unit above a village). The mean village consists of 100 households, and at baseline, the average household had 4.3 members, of which 2.3 were children. The average survey respondent was 48 years old and had about 6 years of schooling. 97% of households were engaged in agriculture; at endline, 49% of households in control villages were also engaged in wage work and 48% in self-employment. Many of the small household enterprises are in petty retail, trade or livestock products. Transfers and data collection took place from mid-2014 to early 2017, a period of steady economic growth, relative prosperity, and political stability in Kenya.

B. *Intervention*

The enrollment of households was relatively inclusive. GD defined as eligible all households that lived in homes with thatched (as opposed to metal) roofs. GD then enrolled all households that met this criterion in villages assigned to treatment. Based on our household census data (described below), 35%-40% of households were eligible. This is far more inclusive than existing public programs in the area, which reached 1.3% of individuals and 6.5% of households in Siaya at the time.[12] That said, the results (described below) may still understate the potential to boost social welfare by targeting even less deprived households.

Eligible households received transfers totaling KES 87,000, or USD 1,871 PPP (USD 1,000 nominal), which constitutes 75 percent of mean annual household expenditure.[13] All transfers were delivered via the mobile money system M-Pesa, and households selected the member they wished to receive them. Transfers were delivered in a series of three tranches: a token transfer of KES 7,000 (USD 151 PPP) sent once a majority of eligible households within the village had completed the enrollment process, followed two months later by the first large installment of KES 40,000 (USD 860 PPP). Six months later (and eight months after the token

[12] Data provided by GiveDirectly, originally from the Government of Kenya's Single Registry for Social Protection.

[13] PPP conversions based on data from World Bank (n.d.).

transfer), the second and final large installment of KES 40,000 was sent. Beyond this point transfers were non-recurring, i.e., no additional financial assistance was provided to recipient households after their third and final installment, and they were informed of this up front. Households in control villages did not receive transfers.

C. *Experimental design and data*

The study employed a two-level randomization design. First, we randomly assigned sublocations (or in some cases, groups of sublocations) to high or low saturation status, resulting in 33 high- and 35 low-saturation groups. Within high (low) saturation groups, we then randomly assigned two-thirds (one-third) of villages to treatment. Randomization was well-balanced with respect to an array of household demographic and economic characteristics (see EHMNW).

We first conducted a baseline household census in all villages, which serves as a sampling frame and classifies household eligibility status. The census was designed to mimic GD's censusing procedure but was conducted by independent (non-GD) enumerators across both treatment and control villages for consistency. The census identified 65,383 households with a total baseline population of 280,000 people in study villages.

Within one to two months after the census, and before the distribution of any transfers to each village, we conducted baseline household surveys. These targeted a representative sample of eight households eligible to receive a transfer and four ineligible households per village. When households contained a married or cohabiting couple, we randomly selected one of the partners as the target survey respondent. We conducted a total of 7,848 baseline household surveys between September 2014 and August 2015, of which 5,123 (66%) were of eligible and 2,722 (34%) of ineligible households, in line with the sampling targets.

We later conducted endline household surveys, targeting all households that had been surveyed at baseline, as well as those that were sampled but missed at baseline, and we attempted to survey the individual who was the baseline respondent. We conducted a total of 8,239 endline household surveys between May 2016 and June 2017, of which 5,423 (66%) were of eligible and 2,816 (34%) of ineligible households. We achieved high respondent tracking rates at endline, reaching over 90% of households in both treatment and control villages, and these rates do not systematically vary by treatment status (Egger et al., 2022*b*, Tables F.1 and F.2).[14]

One valuable property of the endline surveys is that they were conducted over a wide and randomly-assigned range of times relative to the timing of transfers. Specifically, all villages in the study were assigned an "experimental start month" when GD transfers were scheduled to begin if that village were assigned

[14] In addition to household surveys, the study also collected surveys of enterprises, market prices, and local government. EHMNW and Walker (2018) discuss these data and present additional results.

to treatment, and endline surveys were then timed between 9 and 31 months after this date, with the difference also experimentally assigned. Figure A.1 illustrates the resulting distribution of time elapsed between the date when a given shilling was transferred to a household and the date that household's endline survey was conducted. The mode is roughly 13 months, but with substantial mass at both higher values and near a zero time lag (i.e., the household was surveyed in the same month as the final transfer was received). The data are thus informative about predicted deprivation and impact over a relatively wide range of time horizons post-transfer—certainly as compared to PMT exercises that uses covariates to predict contemporaneous deprivation, i.e. with no lag. As we discuss below, the design of the dataset allows us to learn predictive models as a function of time since transfer as well as household covariates.

For the purposes of this paper, we focus primarily on eligible households that were surveyed at both baseline and endline, as we observe them under either treated or control conditions (at endline) and can use baseline values of household characteristics to predict both deprivation and impact. We also require households to have non-missing endline outcome data and baseline covariates.[15] These inclusion conditions yield an analysis sample of 4,749 eligible households. Relative to ineligible households and to the overall population, eligible households tend (as expected) to have lower living standards, as for example measured by the predictions we will obtain for their per-capita consumption (see Figure A.2). But there is still substantial overlap between the groups, indicating that data on impacts among the eligibles do let us examine the relationship between deprivation and impact over a relatively wide range of economic conditions, including both among the very deprived as well as relatively well-off households in these communities.

We use baseline data on a set of 16 covariates (the vector X_h in the framework above) to predict endline outcomes. We selected variables that we found in other real-world proxy means tests used to target social protection problems and that exhibit meaningful variation in our data. The resulting list includes demographic measures (e.g., household size, indicators for children of various ages) and economic measures (e.g., ownership of major assets, employment status); Appendix B.1 provides the full list.[16]

We focus on four pre-specified outcomes at endline, including core household financial outcomes (namely, consumption expenditure, assets, and income) as well as an index of food security. Details of the construction of these aggregates are provided in Appendix B, and in a pre-analysis plan (PAP) that was written

[15] Specifically, we exclude households for which more than 7 baseline covariates were missing (which only drops 3 observations). The Generalized Random Forest (GRF) statistical package (discussed below) handles missing covariate values by considering the missing status itself as a potential split on that variable, allowing missing values to be informative.

[16] As discussed in detail in the Appendix, we select predictors by hand rather than using the specific data-driven approach we had originally pre-specified, as the latter was not well-defined and creates issues for inference. That said, the main results are all qualitatively robust to using a data-driven approach instead.

specifically for the targeting analysis that is the focus of this paper and that is posted on the AEA RCT Registry (at `https://www.socialscienceregistry.org/trials/505`). In the main analysis, we predict versions of these outcomes demeaned by the month in which the survey was conducted, in order to remove any effects due purely to correlation between predictors and survey timing, and then add back in the overall mean to all observations for interpretability. Demeaning by survey month is important since some households are easier to contact than others, potentially resulting in baseline characteristics being predictive of survey timing even though timing at the village level was randomized.

The three financial outcomes—consumption expenditure, income, and assets—are defined at the household level, the same level at which treatment was assigned, so that they correctly capture the total effects of treatment as opposed to their per-capita analogues (which would under-weight impacts on individuals living in large households). Recall that cash transfers of the same magnitude were provided to all treatment households regardless of the number of members. Taken together, these outcomes form a natural constellation given their connection via the household's budget constraint, and studying them in tandem allows us to relate the results to canonical dynamic models of consumption and investment. For example, if households vary in their marginal propensity to consume (MPC) as opposed to investing out of a transfer, then we would expect to see negative covariation between initial treatment impacts on consumption and accumulated assets. Over time, however, the households that invested more should realize higher levels of income, consumption, and assets. If households vary mainly in the returns their investments yield then we might see positive covariation emerge quickly, especially since (as we discuss further below) typical investments in this setting would likely have very short gestation periods.

Food security is an important public policy objective for many transfer programs (though these are usually structured as streams of small payments, as opposed to the lump sum transfers studied here). It is also theoretically interesting as a case in which we might expect *a priori* to observe a relatively weak tradeoff between targeting on deprivation versus impact, given that the households most likely to spend on better nutrition are often those not eating enough (see for example Subramanian and Deaton, 1996). Unlike the total household financial outcomes noted above, the index of food security we use is arguably best interpreted as a per capita measure: typical constituent questions ask how many days (out of the past 7) family members experienced a negative outcome such as skipping meals, a quantity we would not expect to scale mechanically with household size (as for example total household food consumption would). Indeed, we will show below that results for the food security index parallel those for per capita food consumption, and that these both differ from results using total household food consumption.

D. *Existing results*

EHMNW report the overall average impacts of the GD program on recipient households, estimating positive ITT effects on each of the four outcomes we consider here, among others. They also find large spillovers onto untreated households, for example, substantial expenditure increases for non-recipient households and higher enterprise revenue in areas that received more cash transfers. Using these and related estimates, they derive the implied multiplier effect on overall local economic activity, estimating a transfer multiplier of 2.5.

Given these spillover results, the analysis that follows should be interpreted as examining variation in who is selected for treatment, holding fixed the total number of local households treated. Spillover effects do not alter this analysis to the extent that they are approximately additive and invariant to the identity of the household receiving the transfer or experiencing the spillover. We cannot readily estimate the extent to which different kinds of households generate different spillovers; as we discuss below, this would require an experiment even larger than our (already very large) one. We can, however, use several complementary strategies to assess the extent to which different kinds of households are affected differently by spillovers; we return to this issue below.

With respect to heterogeneity of treatment effects, EHMNW take the conventional approach of testing across a pre-specified, researcher-selected set of covariates (including, for example, respondent gender, age, marital status, and educational attainment, among others). They generally fail to reject homogeneity of treatment effects along these dimensions but are only moderately powered to detect effects (Table E.1, derived from data in EHMNW). We therefore turn next to examining data-driven ML approaches to identifying features of the baseline data that (potentially) predict deprivation and impact.

III. Empirical methods

This section describes the empirical methods used to operationalize the ideas outlined in the conceptual framework. Broadly speaking, the approach is to (i) predict (per capita) outcomes absent treatment, and treatment effects, for each household as a function of its baseline covariates (X_h) and time since treatment t; (ii) integrate over t to obtain time-averaged predictions; and then (iii) classify households into groups based on whether they are or are not selected to receive transfers under various social welfare functions (including the limit cases in which the planner targets solely impact or solely deprivation). We then measure deprivation and impact within the groups selected by this procedure using simple OLS estimators. We discuss among other things the approach to regularization and to inference. The analysis follows a pre-analysis plan specific to this targeting analysis, submitted to the AEA registry on 1 September 2017, prior to the estimation of treatment effects for these outcomes.[17]

[17] See `https://www.socialscienceregistry.org/trials/505`.

We work in particular with the constant absolute risk aversion (CARA) function

$$W(\hat{y}) = \begin{cases} (1 - e^{-\alpha\hat{y}})/\alpha & \alpha \neq 0 \\ \hat{y} & \alpha = 0 \end{cases} \tag{10}$$

which is commonly used in applied work. We consider a range of values for the curvature parameter α, starting from $\alpha = 0$ (the linear case corresponding to targeting based solely on impact) and gradually increasing from there up to 0.015. This range includes most of the estimates in the literature reviews by Barseghyan et al. (2018) and Babcock, Choi and Feinerman (1993). We note that an α value of 0.015 implies an extreme amount of redistribution in our setting, as the planner would value a marginal dollar to the person at half the average level of consumption (in our sample) at over 360 times as much as a marginal dollar to the person at the average. Finally, we consider as a benchmark the limit case $\alpha \to \infty$ which corresponds to targeting based solely on deprivation, the approach most often studied in the literature and the focus of public policy.[18]

The curvature of the utility index function $W(\hat{y})$ defines the relative value of the marginal dollar, which in turn characterizes how much the policy maker up-weights the contributions of transfers to the poorest households relative to transfers to the richest households in the social welfare maximization. While CARA utility is sensitive to scale, and therefore the units of the outcome, the range of parameter α values that we consider encompass all reasonable levels of curvature given the units and magnitudes of our primary outcomes. These levels of curvature could be generated by alternative functional forms. For instance, our range of α parameters being considered imply relative risk aversion levels (a unit-less measure of curvature) in the range of $\rho \in [0.0, 11.8]$ for the mean consumption per capita in our sample. With an alternative functional form for the utility index such as the Constant Relative Risk Aversion (CRRA) function, a relative risk aversion value of $\rho = 11.8$ would imply that the policy maker weights a marginal dollar to the person at half the average level of consumption (in our sample) at over 3500 times as much as a marginal dollar to the person at the average.

As an alternative way to introduce curvature to our welfare analysis, we also examine the sensitivity of our conclusions to using a Constant Relative Risk Aversion (CRRA) function. This raises a technical difficulty as in a small minority of cases the predicted per-capita outcomes from our model are negative, so that CRRA is undefined. We deal with this by shifting the distribution of per capita

[18] Estimates based on private risk-taking decisions are also available from a setting close to ours, the Busara Center lab in Nairobi, where Balakrishnan, Haushofer and Jakiela (2020) estimate average values of about 0.0013 for Kenyan shillings, which corresponds to 0.052 for US dollars. These estimates imply a high degree of risk aversion that falls on the boundary of the estimates reported in the literature, e.g. by Babcock, Choi and Feinerman (1993). Balakrishnan, Haushofer and Jakiela (2020) obtained these estimates using stakes corresponding to about 4 times the median daily expenditure. Because concavity is typically stronger over smaller stakes (Rabin, 2000), it is likely that they are overestimates relative to the degree of concavity one would observe over policy-relevant stakes.

consumption so that its minimum value equals one-fourth of the World Bank's extreme poverty line, on the rationale that it would be difficult to survive on less than that for any length of time, and obtain qualitatively similar results when we do so for parameter values that generate similar relative weights to those implied by our CARA estimates. In particular, we consider relative risk aversion coefficients in the $[0, 4]$ range. This range of parameters includes the mean estimate in a literature review of 92 studies by Elminejad, Havranek and Irsova (2022), and includes the upper bound of values that rationalize established facts about labor markets found by Chetty (2006) using estimates from various settings and countries. Given that this procedure is inherently somewhat arbitrary, however, the CARA estimates are our preferred ones.[19]

Because the outcomes in the data are measured at the household and not the individual level, the analysis needs to account for variation in household size. Generalizing Equation 2 by interpreting Y_h as a household aggregate, denoting by n_h the size of household h, and weighting all household members equally, the planner's objective function is

$$\sum_h \sum_{t=0}^{\bar{t}} n_h W(Y_{h,t}(T_h)/n_h) = \sum_h \sum_{t=0}^{\bar{t}} n_h W(Y^0_{h,t}/n_h + T_h \cdot \Delta_{h,t}/n_h) \tag{11}$$

Note that in this empirical setting the size of the transfers (and thus the cost of treatment) are the same irrespective of household size. We would therefore expect per capita treatment effects to be mechanically smaller in larger households, but this does not mean that they are less attractive to target. Indeed the precise details of optimal targeting here depend on the interplay of the joint distribution of $(n_h, Y^0_{h,t}, \Delta_{h,t})$ with the curvature of W, something that is captured in the welfare analysis. That said, the planner generally prefers to target households with large absolute treatment effects $\Delta_{h,t}$ and with low per capita outcomes absent treatment (denoted henceforth by $y^0_{h,t} \equiv Y^0_{h,t}/n_h$).[20] We therefore begin the analysis by identifying the households predicted to be most deprived on a per capita basis, and those most impacted on an absolute basis.[21]

At the core of this approach is the classification procedure summarized in Algorithm 1 (and in greater detail in Algorithm E.1). The procedure predicts the full joint distribution of deprivation and impact, and then uses these predictions to

[19] Along with CARA and CRRA preferences we also prespecified the inequality-averse preferences of Fehr and Schmidt (1999), but these depend on pairwise comparisons that turn out to be computationally prohibitive in our setting, and in any case are not widely used for social welfare analysis in the literature.

[20] To see this, note that for small treatment effects welfare is well-approximated by the first-order expansion

$$\sum_h \sum_{t=0}^{\bar{t}} n_h W(y^0_{h,t}) + \sum_h \sum_{t=0}^{\bar{t}} W'(y^0_{h,t}) \cdot [\Delta_{h,t} \cdot T_h] \tag{12}$$

so that the incremental benefit at time t of treating h is approximately $W'(y^0_{h,t}) \cdot \Delta_{h,t}$.

[21] We abstract from issues of intra-household inequality, which the data do not let us examine.

```
Algorithm 1: Select most-deprived and most-impacted groups
Split data into set $\mathcal{K}$ of folds;
foreach $K \in \mathcal{K}$ do
    Training data $K' \leftarrow \mathcal{K} \setminus K$ other folds ;
    $\{\hat{y}^{0,K} : \{\mathbf{X}, T\} \to \mathbb{R}\} \leftarrow$ predictor of $y^0_{h,t}$ learned from training data $K'$;
    $\hat{y}^{0,K}_h \leftarrow \frac{1}{\bar{t}} \sum_{t=0}^{\bar{t}} \hat{y}^{0,K}(X_h, t)$, i.e. integrate over time;
    $\{\hat{\Delta}^K : \{\mathbf{X}, T\} \to \mathbb{R}\} \leftarrow$ predictor of $\Delta_{h,t}$ learned from training data $K'$;
    $\hat{\Delta}^K_h \leftarrow \frac{1}{\bar{t}} \sum_{t=0}^{\bar{t}} \hat{\Delta}^K(X_h, t)$, i.e. integrate over time;
    Classify observations in top 50% of $\{dW(\hat{y}^{0,K}_h, \hat{\Delta}^K_h)\}$, $h \in K$, as socially
      optimal given $W$;
end
```

classify every household in the dataset as either in or out of the set of households that would be selected for treatment under a given function W.[22] This procedure aims to reduce the risk of over-fitting by classifying each observation h into groups without making any use of its own outcome $Y_{h,t}$; h is instead classified using a function learned only from folds of the data that do not include it. We set $K = 5$, and (to ensure results are not sensitive to the specific split into K folds) then repeat the entire procedure 150 times and report mean outcomes across these iterations.[23,24]

Predictions are formed by learning the regression function $\mathbb{E}[y^0_{h,t}|X_h, t]$ through random forests and the conditional average treatment effect (CATE) function $\mathbb{E}[Y^1_{h,t} - Y^0_{h,t}|X_h, t]$ through causal forests, using the generalized random forests (GRF) package of Athey, Tibshirani and Wager (2019). We pre-specified an approach based on random forests as these are an attractive tool for uncovering heterogeneity in this setting.[25] Specifically, the dimensionality of our predictors

[22] In practice, we learn models for endline per capita values using the full dataset (i.e., including both treated and control individuals) while including an indicator for treatment status among the predictors. Results are similar if the model is trained on control group data only (Tables D.1, Panel B and D.3).

[23] Classification thresholds are defined for each fold using only their predictions to avoid overfitting concerns since these are not trained using that fold's data. Therefore, a higher number of folds leads to fewer data points being used to define these thresholds. On the other hand, a lower number of folds leads to fewer data points being used to train each random forest. Given our sample size, 5 folds leads to reasonable subsample sizes for each of these steps. Note also that while we use common splits to learn $\hat{y}^0$ and $\hat{\Delta}$, we obtain essentially identical results if we use separate splits.

[24] An alternative to computing $W(\hat{y})$ is to first calculate $W(y)$ and then learn models to form predictions $\widehat{W(y)}$ directly, as in the Empirical Welfare Maximization literature. Empirically we find that learning models perform relatively poorly on the transformed $W(y)$ data due to the wide range of numeric values they take on, however. Our application differs in this regard from the empirical application in Kitagawa and Tetenov (2018), for example, who consider maximization of the average treatment effect on (untransformed) earnings and in a setting where baseline household income is much higher than in ours.

[25] The pre-analysis plan specified that we would implement the causal forests approach of Wager and Athey (2018) or methods that improved on it, if any were available by the time data were collected. We therefore implement Athey, Tibshirani and Wager (2019) which generalizes and extends Wager and

is low relative to the number of observations and we do not see strong evidence of heterogeneity along dimensions that we (originally) thought might matter. Random forests are particularly well-suited for dealing with such non-sparse settings, and can account for complex non-linearities and interactions between the predictors.

At the same time, using a regularized method is important in an optimal targeting context to mitigate the risk of over-fitting. Naive methods—based for example on OLS—might claim to identify very deprived households or those with large treatment effects, leading to over-stated estimates of the overall anti-poverty impact of a program or to mis-estimation of the tradeoff between deprivation and impact. Regularized methods such as random forests help to address this risk.[26] We report forest-based results as our preferred estimates, and also benchmark these against results using OLS and alternative ML estimators in Section V.C.

To calculate time-averaged predictions we first learn predictive models using the number of months t since the "experimental start month" in each village as a predictor. For each observed value of X_h, we then evaluate these models at 7 quarterly intervals, i.e. over a total range of 21 months. Finally, we take an unweighted average of these predictions to obtain time-averaged predictions.

Given a classification of the sample into groups indexed by S, we define the following measures of performance. The **predicted averages** are the within-group means of GRF predicted values:

$$\bar{\hat{y}}^0(S) = \frac{1}{|S|}\sum_{h\in S}\hat{y}^0(X_h) \qquad \bar{\hat{\Delta}}(S) = \frac{1}{|S|}\sum_{h\in S}\hat{\Delta}(X_h) \tag{13}$$

These may or may not be consistent for the results a policymaker would actually obtain by targeting group S. While our procedure guards against over-fitting in forming predictions $\hat{Y}_h^0/n_h$ and $\hat{\Delta}_h$ for individual households, targeting requires us to take the additional step of selecting groups of households based on these predictions. This introduces the additional risk of a "winner's curse." To the extent there is even non-systematic error in the predictions, we will tend to select observations with extreme values of this error. For example, we will tend to classify households with high values of $Y_h^0 - \hat{Y}_h^0$ as deprived, and thus to over-estimate how deprived the most deprived group is.

To address this issue, we also calculate a separate set of **actual averages**

Athey (2018). In parallel Chernozhukov et al. (2018) developed attractive methods for learning average treatment effects and characterizing units within quantiles of the treatment effect distribution; for our purpose here, however, we require the unit-level predictions that GRF provides.

[26] The GRF package in particular uses cross-fitting and an "honest" approach to growing trees to control over-fitting, and we add to this by classifying each observation without using data from its own fold. Random forests do require some tuning and, unlike for other ML procedures such as LASSO, optimal regularization procedures are not available. We selected tuning parameters from among two options: the GRF package defaults, and an alternative set suggested by one of the authors of the package as a way to provide stronger regularization (see `https://github.com/grf-labs/grf/issues/120#issuecomment-327276697`, accessed 31 August 2021). We use the latter as it provides a closer match between predicted and actual statistics.

which are simply group means (for y^0) or group average treatment effects (for Δ) estimated via OLS:

$$\overline{y}^0(S) = \frac{1}{|S|}\sum_{h\in S} y_h^0 \qquad \overline{\Delta}(S) = \frac{2}{|S|}\sum_{h\in S}\left(Y_h^1 T_h - Y_h^0(1-T_h)\right) \tag{14}$$

This approach uses predictions of deprivation $\overline{\hat{y}}^0(S)$ and impact $\overline{\hat{\Delta}}(S)$ only to select groups, not to estimate outcomes within those groups. We interpret the comparison between predicated and actual averages as a measure of how successfully our approach predicts results in these groups, where smaller gaps are indicative of better performance. We expect this comparison to be an especially stringent test when we conduct it for the groups estimated to be most deprived (D) and most impacted (I), since in those cases the risk of a "winner's curse" bias is most pronounced.

For inference we report bias-corrected and accelerated (BCa) confidence intervals based on bootstrapped values of statistics (Diciccio and Romano, 1988; DiCiccio and Efron, 1996). These have the advantage that they can be asymmetric and adjust for any potential skewness in the bootstrap distribution, reflecting the potential asymmetry involved in selecting maximal elements from a set of statistics.[27] For the main results we bootstrap the entire procedure including both classification of households into groups and prediction or estimation; for all results, including main results and robustness checks, we report confidence intervals from a bootstrap conditional on household classification (which is far cheaper computationally).

Diagnostics suggest that the procedure, and in particular the repeated 5-fold splitting, produces fairly stable results. Figure E.1 shows, for example, that the mean differences between treatment effects in the most deprived and most impacted groups remain more or less constant if we increase the number of splits from 150 to 300.[28] Figure E.2 shows that the classification of households into most deprived and most impacted groups are also quite stable, with most households assigned fairly consistently to either one or the other group.

IV. Results

This section presents the main results of the social welfare analysis. Section IV.A identifies the group a policymaker would target given a particular social welfare function, how this group overlaps with the most deprived and the most impacted groups, and how these groups compare in terms of the average levels of deprivation and impact within them. Section IV.B examines the sensitivity

[27] GRF provides asymptotic inference for individual predictions $\hat{y}_h^0$ and $\hat{\Delta}_h$ but not for their joint distribution, so approaches like that proposed by Andrews, Kitagawa and McCloskey (2021) are not available.

[28] We nevertheless report results for 150 splits, since bootstrapping statistics is very computationally costly at 150 splits and would be yet more so at 300.

of target group selection to a number of alternative specifications of the social welfare function, and Section IV.C discusses some broader implications of the results for modelling poverty dynamics.

A. *Optimal policy under concave social welfare functions*

We begin by visualizing the joint distribution of predicted deprivation (absent treatment) and predicted treatment effects along with locally smoothed regression fits (panels A-C of Figure VI). The upper x-axis denotes the quantiles of the distribution of predicted deprivation, while the lower x-axis reports the corresponding monetary values (in USD PPP). One noticeable feature of the distributions for all three outcomes is that there is substantial variation in predicted impact conditional on predicted deprivation, and vice versa. Even absent any systematic relationship between impact and deprivation, this variation—assuming that it is supported by diagnostic checks for over-fitting (see below)—creates a trade-off between the two: some households happen to be high-impact and low-deprivation, while others happen to be low-impact and high-deprivation, and the planner must prioritize between these.

There is also some evidence of systematic covariation between deprivation and impact, particularly for consumption and assets. Here the slope of the non-parametric fit is positive, indicating that less-deprived households also tend to see larger gains when treated. Income displays a slight positive relationship over most of its range, albeit more muted. Panels D, E and F show that these relationships are stable over time, plotting local regression fits as in the top panels but broken down by quarter since cash transfer receipt. Treatment effects fall somewhat on average over time but are consistently positively correlated with untreated outcomes, implying a persistent tradeoff between deprivation and impact. The potential trade-off between short-run investment and consumption is one reason that it is important for us to be able examine the dynamics of consumption (from 1 to 7 quarters after cash transfer receipt); this enables a richer analysis than previous studies, many of which have been largely static.

Note that this pattern is the opposite of what one would expect if the algorithm were over-fitting variation in the data that resulted from classical measurement error, or mere sampling variation. In those scenarios, types of households that happen by chance to have low (high) reported values when untreated would be predicted to be both more deprived and more impacted (less deprived and less impacted). Indeed, we will see precisely this pattern below when we examine results obtained using OLS (rather than GRF), which is prone to over-fitting noise in the data. To further investigate the potential influence of measurement error, we also examine results for a subset of assets that would have been observable to our enumerators as they conducted surveys—such as solar panels or large pieces of furniture, for example—and are thus likely measured with less error. We obtain qualitatively similar results using this subset (Tables A.3 and A.4).

Targeting using a concave social welfare function amounts to passing a curve through the points in Figure VI, selecting those that are above and to the left of it for the program. As benchmarks, the coloring in Figure VI illustrates the extreme cases of targeting solely based on deprivation or on impact. To illustrate the selection process for a social welfare function that trades these two objectives off against other more continuously, Figure 2 plots the groups selected using a CARA social welfare function defined for a range of curvature values ($\alpha = 0, 0.001, 0.015$), where the lower end of the range contains empirically observed values and we view the higher end as extreme degrees of curvature (as noted above). Note that for clarity of illustration this figure presents one specific cut of the data, with the x-axis scaled in USD rather than quantiles as in Figure VI.[29] The middle panel shows that for a "central" curvature value $\alpha = 0.001$ (in line with estimates in the literature) the frontier between the groups selected versus not slopes upward; the selected group includes both some households that are among the most impacted but not the most deprived, and also some that are among the most deprived but not the most impacted. In the linear case $\alpha = 0$ only those who are most impacted are selected (left panel), and for an extremely high value $\alpha = 0.015$ the group selected coincides almost exactly with the most deprived (right panel).

Table 1 describes group selection more comprehensively using the full dataset and all folds of the data. Specifically, it characterizes the groups selected under various social welfare criteria in terms of their overlap with the most deprived group (Column 2) and the most impacted group (Column 4). Columns 3 and 5, respectively, report bootstrap-based p-values from a test of the null hypothesis that these overlaps are at least 95% (a test of the null of 100% overlap is arguably too easy to reject, as a single data point can disprove it).[30]

A failure to reject the null tested in Column 2 indicates that we cannot reject that the current real-world practice of solely targeting the most deprived households is socially optimal.

Considering first the results for consumption (Panel A), we see that as expected the planner selects 100% of the most impacted group when there is no curvature in the SWF (i.e. $\alpha = 0$). As curvature increases, they select a smaller share of the most impacted and a larger share of the most deprived, again as expected. We reject the null of (near-)complete overlap with the most impacted group, signifying that targeting deprivation does matter at these levels of curvature. But we also reject the null of (near-)complete overlap with the most deprived group for any plausible curvature value. Figure VI presents this pattern visually and for a wide range of curvature values; it is not until we reach extreme levels of curvature (approximately $\alpha > 0.01$) that 0.95 lies within a 95% confidence interval for the overlap between the socially optimal group and the most deprived. Recall that at

[29] Specifically, it illustrates predicted values for consumption for households of median size (i.e., 4 members), using a static model, with the thresholds that define groups obtained without cross-fitting and held constant across the whole sample.

[30] Appendix Table A.11 shows the sensitivity of this test to different thresholds; at a null of 100% overlap it almost always rejects, while for less extreme values it rejects as expected depending on the value of α.

$\alpha = 0.015$—the highest value we consider and a value at which nearly all deprived (D) households would be targeted—the planner values a marginal dollar to the person at half the average level of consumption (in our sample) at over 360 times as much as a marginal dollar to the person at the average. We view this as an empirically implausible level of redistributive social preferences.

As an alternative way to quantify how extreme social preferences would need to be to rationalize targeting only the most deprived, one can also back out the ad hoc Pareto weights a planner would need to place on that group. With a moderate degree of curvature in the CARA SWF (at $\alpha = 0.0005$), the planner would need to weigh the deprived 3.76 more than the non-deprived to justify fully targeting the D group, above and beyond the weighting implied by the curvature of the social welfare function itself. We view this degree of additional redistributive preferences as empirically very large.

Results for assets and income are similar. In the interests of brevity results are reported here (Panel B of Table 1) for only a central curvature value ($\alpha = 0.001$), with results for a wider range in Table A.1. As for consumption, the planner selects (nearly) all of the most impacted only at very low levels of curvature, and (nearly) all of the most deprived only at extremely high levels. At plausible levels of curvature, both deprivation and impact matter and many of the most deprived households are not optimally targeted. This is a central finding of this study.

The results indicate that different welfare criteria select very different groups; how different are those groups in terms of their levels of deprivation and impact? The visualizations in Figure VI suggest that the differences are meaningful. Table 2 examines this more systematically, reporting mean levels of deprivation and impact for the most deprived and most impacted reference groups to illustrate the trade-offs that the optimization is balancing, as well as values for the socially optimal group it ultimately selects. We focus on the actual values of these statistics as defined above, but also report their predicted values in order to compare the two as a diagnostic check on the predictive validity of the methods.

There is a quantitatively meaningful trade-off between deprivation and impact. For all three outcomes the average outcome among the most deprived (Column 2) is substantially lower than the overall average (Column 1)—by 31%, 75%, and 43% for per capita consumption, assets, and income, respectively. Evidently the predictors contain enough information to identify a sub-population substantially more deprived than average, even among a population that has already been selected to be poorer than average using GD's coarser targeting criterion. Targeting the most impacted, on the other hand, comes at a substantial cost in terms of targeting deprivation. Column 4 reports endline values in the absence of treatment for the group identified by the model as most impacted by treatment. In contrast to the most deprived group, the most impacted group is actually better-off than average for each outcome. Relative to the overall sample mean, their levels of per capita consumption, assets, and income are higher by 25%, 58%, and 9%, respectively. As a result, the differences in deprivation between the most deprived and most impacted groups are also large and statistically significant (Column 5)—consumption, for example, is 45% lower among the most deprived than among the most impacted. As expected, deprivation in the socially optimal group (Column 3, evaluated at $\alpha = 0.001$) falls in between levels for the most deprived and most impacted; notably, it has a lower mean than the overall average.

Turning to impact (Columns 6-10), we see the other side of the trade-off. Impacts in the most deprived group (Column 7) are consistently below the overall average treatment effect (Column 6). In contrast, outcomes for the most impacted are (as expected) consistently above average (Column 8). The net result is that targeting the most impacted as opposed to the most deprived yields substantial gains in treatment effect. Treatment effects are 64%, 22%, and 19% larger in the most impacted group relative to the most deprived group for consumption, assets, and income, respectively (Column 10), with these differences statistically significant at the 0.07 level or lower for all three outcomes. Predicted values lie between these two extremes for all outcomes, while actual values for consumption and income are in fact slightly higher than in the most impacted group. The considerable variation in predicted impacts is consistent with the existing finding of highly heterogeneous returns to investment in many low- and middle-income country settings (see Demirgüç-Kunt et al., 2022).[31]

Finally, we note that actual outcomes line up quite closely with those predicted by the model. The most deprived are in fact *more* deprived than the procedure predicted them to be across all three outcomes (Column 2). The most impacted are slightly less impacted than predicted in the case of income, but substantially more impacted than predicted in the case of consumption and just as predicted in the case of assets. Overall these comparisons suggest that the regularization and cross-fitting procedures built into the selection procedure are largely effective at mitigating over-fitting and "winner's curse" effects, since these would tend to lead to over-optimistic predictions about the extremes of deprivation and impact we can identify. We emphasize that these results are important for the interpretation of the earlier group selection results; if the diagnostics in Table 2 had performed poorly, this would have indicated that the hypothesis tests in Table 1 were likely to over-reject the null because the algorithm was "detecting" heterogeneity that was in fact noise.

B. Alternative social welfare criteria

The core conclusion from optimization using the baseline CARA social welfare function is that for plausible levels of curvature, targeting should reflect both deprivation and impact rather than just deprivation (or impact) alone. We next examine how sensitive this conclusion is to variations on the social welfare function, with results reported in Panel C of Table 1. We focus on consumption

[31] Comparisons across outcomes in Table 2 are not straightforward since the groups selected are themselves different for each outcome. Similar conclusions hold, however, if we select groups based on their consumption and then examine values for all three outcomes—see Appendix Table A.9.

throughout.

The first two variations consider sensitivity to functional form, replacing the CARA utility function with a CRRA function, $W(y) = y^{1-\rho}/(1-\rho)$, evaluated at two different levels of curvature, namely, $\rho \in \{0.5, 2\}$. Note that because a handful of predicted values are negative (due to the time de-meaning and re-centering steps in the algorithm) we shift the outcome distribution so that the minimum value is 200 USD at 2017 PPP, which is roughly 1/4 of the World Bank's poverty line (World Bank, 2022) and an amount below which it seems unlikely one could survive for any length of time. The third considers sensitivity to time discounting, augmenting the base CARA model (with $\alpha = 0.001$) with an aggressive discount rate of 15% per annum. In each of these cases the main conclusions are unchanged, and point estimates are similar to those in the base scenario.

The fourth variation considers a different type of dynamics, asking how a *future* round of re-targeting might lead to different outcomes if we first implemented the targeting rule implied by the baseline analysis today. To examine this we select the group that is socially optimal under CARA with $\alpha = 0.001$ and add each member's predicted treatment effect to its predicted untreated outcome to obtain a new baseline level of consumption. We leave predicted consumption levels unchanged for households not selected for treatment. This yields a new joint distribution of treatment effects (which is unchanged) and deprivation (which has changed, since some households were previously treated). We then re-select recipients a second time by applying the same welfare maximization procedure to this new joint distribution.

The analysis indicates that the composition of the groups selected changes non-trivially: around 12% of the households targeted in the first round are not targeted in the second round (see Table A.10). The share of targeted households that were initially among the most deprived, however, is approximately the same, at 53% (as opposed to 54% in the first round). This is because the first round of transfers lowers the marginal utility of treated households in the most-deprived group as well as those in its complement, so that some most-deprived households become more attractive to target while others become less. At extreme levels of curvature the latter effect dominates, so that substantially fewer of the initially most deprived are targeted in the second round than in the first (78% as opposed to 96%; see Table A.10, right-hand panel). An overall takeaway is that additional targeting iterations do not necessarily tilt the balance back towards those who were initially most deprived.

The fifth variation adds Pareto weights to equation (3) to capture the idea that society may want to consider factors other than consumption when determining who is deserving of help. There are, of course, arbitrarily many weightings for which one could argue. For concreteness we take advantage of data from Alatas et al. (2012), who elicited community rankings which were used to allocate cash transfers in Indonesian villages. In addition to these rankings the data also contain

contain consumption, household size, and two covariates—widow status and the presence of children—that overlap with our predictors. As a result we can use their data to estimate Pareto weights as a function of those overlapping predictors, as well as the curvature parameter α, and then apply these parameter estimates to optimization on our own data. Appendix G provides the details of this procedure.

The results indicate that the Indonesian data imply a sizeable 47%-60% higher weight for households headed by widows, a more modest 17%-29% higher weight for households with children than for those without, and a curvature parameter of $\alpha = 0.0004$. We take the estimated widow weight into our setting in Kenya (weights on having children were only marginally significant and focusing on one dimension allows us to better isolate the role of these weights). When setting $\alpha = 0.0005$ (very close to the Indonesia estimate, and facilitating comparison with the baseline estimates in Table 1, Panel A), we see that this upweighting does modestly increase the share of the most deprived households that are socially optimal to target, as expected. Yet we can still reject that the planner would prefer to only target the most deprived even in this case ($p < 0.01$).

We also consider the alternative approach to welfare analysis suggested by Saez and Stantcheva (2016) in which *all* the distributional considerations are embedded into "generalized social marginal welfare weights." As there is no general theory or principles from which these weights should be derived, we again work with the data from Alatas et al. (2012) to pin them down. Relative to the Pareto weights exercise the key difference lies in how we apply these estimates to the Kenyan data; now (i) weights depend on household per capita consumption as well as on other characteristics, but (ii) households are ranked for inclusion according to their weighted treatment effect, without applying any other curved utility function (i.e. effectively $W(y) = y$). Notice that in this approach the same factors that mattered above—household size, per capita income, and the other characteristics in X_h—are still allowed to influence the allocation, but their influence now works solely through the marginal weights. All that said, the group selection we obtain in the end is indistinguishable from the Pareto weights approach, and more generally about the same as in the baseline approach (Columns 2 and 4).

All told, these exercises illustrate a number of fruitful ways in which the baseline social welfare analysis can be adjusted or extended. At the same time, they also demonstrate that the basic conclusion—that welfare-maximizing targeting should reflect both deprivation and impact, rather than deprivation alone, as is commonly the case in real-world practice today—is robust to a wide variety of welfare criteria.

C. *Economic interpretation*

The results above are of natural interest from the point of view of theories of poverty dynamics, as any such theory will yield predictions, implicitly or explicitly, about the joint distribution of deprivation and impact. We next summarize five descriptive facts that appear particularly relevant for thinking about this mapping.

First, household size is, by a wide margin, the most important predictor of both deprivation and impact. We can see this in Table 3, which summarizes the predictive importance of each of the 16 elements of X_h. We measure importance here (as does the GRF package) as a depth-weighted average of the share of splits created in the process of growing trees.[32] A value of 0.05 for "female head," for example, means that when growing trees the algorithm chose to split on whether or not the household had a female head in 5% of cases. Numbers in parentheses indicate the rank of each predictor's importance within that column, and the signs indicate whether it predicts the outcome positively or negatively. The three most important predictors in each column are indicated in bold. In all six cases household size is ordinally the most important, and cardinally far more so than the next-most-important predictor.

This pattern is not mechanical. Transfers are fixed irrespective of household size, so there is no a priori reason to expect larger effects in larger households. As for deprivation, household size is in the denominator of y_h^0 by construction, so that any measurement error will tend to induce a negative relationship—yet larger households still have noticeably higher per-capita values. These patterns call to mind the classic idea of scale economies in household production (Nelson, 1988; Deaton and Paxson, 1998), or of risk diversification, as households with more members may be better able to spare one to undertake risky, higher-return ventures. Consistent with this idea, the most impacted households have substantially more working-age adult members than do the most deprived across all primary outcomes (Figure A.3, Column 1).[33] Note that this is the number of working-age adult members in the household, rather than their proportion; households with more working-age adults tend to also have more children present.

Second, the tradeoff between deprivation and impact appears to be related to life-cycle dynamics. One clue to this is in Table 3, where the second-most important predictor of deprivation is "having an elderly member." For a more thorough examination, Column 2 of Figure A.3 plots the distribution of the age of the household head separately for each group. The most deprived are disproportionately likely to be either young or old, while the most impacted are more likely to be either young or middle-aged adults. The issues of how much to emphasize targeting deprivation as opposed to impact is thus related to the issue of what stage in the life-cycle to target (calling to mind, for example, debates about

[32] The formula is

$$\text{Importance}(x_j) = \Sigma_{k=1}^{4} \left[\frac{\Sigma_{\text{all trees}} \text{ number depth } k \text{ splits on } x_j}{\Sigma_{\text{all trees}} \text{ total number depth } k \text{ splits}} \right] \Big/ \Sigma_{k=1}^{4} k^{-2} \tag{15}$$

Note that this metric sums to 1 across all covariates in the model.

[33] In an interesting contrast, land ownership is not a strong predictor of deprivation or impact. This is partly because it simply does not vary greatly (with 85% of households owning land), but likely also because—unlike in some other agrarian settings—non-land holders in our context are likely to be profitably engaged in commerce or non-agricultural employment as opposed to working on other people's farms.

whether to allocate scarce social protection resources to parents of young children or to the elderly via old-age pensions).

Third, larger impacts on financial outcomes do not appear to come at an opportunity cost on other, less-salient margins. In particular, one might worry that high-impact households simply decrease their leisure hours more, or reduce their (net) transfers to other households more, to achieve these gains. Tables A.5, A.6, and A.7 report differences in treatment effects (and baseline values) for these outcomes, as well as many others, between the most deprived and most impacted groups. Whether classifying households based on consumption, income, or assets, we see no significant differences in transfers sent or loans given. (If anything the most impacted households see a modest increase in transfers received, which seems more consistent with crowding-in resources in response to new market opportunities.) Impacts on hours worked are similarly not significantly different, and in two out of three cases the difference is actually negative.

Fourth, there is some suggestive evidence that credit constraints play a role. The most impacted households are also *both* more likely to have received a loan and more likely to have been denied a loan in the last 12 months, which would be consistent with greater demand for credit and credit constraints for this group at baseline (Tables A.5-A.7). Along with the household size results, this pattern seems broadly consistent with the idea that some households are better situated to take advantage of the new opportunities that transfers afford.

Fifth, and related, the same households tend to see larger effects on all three financial outcomes (consumption, income, and assets). Figure 4 illustrates this, presenting the marginal densities of predicted treatment effects on each outcome (top row) followed by scatterplots of the pairwise joint distributions of effects on two outcomes at a time (middle row), with correlation coefficients indicated. We see that predicted treatment effects on any one outcome are very strongly correlated ($r \in [0.33, 0.65]$) with those on either of the other two. In part this is, of course, a necessary consequence of the fact that household size is such a strong, common predictor of impacts on all three outcomes.

It is natural to interpret this pattern through the lens of standard consumption-savings frameworks. To the extent households differ either in their initial propensity to invest, or in the returns they make on those investments, we would expect to see positive covariation between all three financial outcomes emerge over time. One key difference between these two mechanisms, however, is in what happens in the early days immediately after transfer receipt. If differences in the marginal propensity to invest were the main driver, we would expect to see much a much more negative relationship between consumption and assets in those early days and months. In the data, however, the pattern is if anything the opposite: the correlation between effects is quite flat over time, and if anything slightly higher in earlier quarters (Figure 4, bottom row). This suggests that differences in the returns households realize is the more important source of heterogeneity. This would be consistent with the types of investments households in this context commonly make. For instance, many small retail businesses purchase inventory that is sold immediately, and those who purchase livestock will soon have eggs and milk to sell. Even those who expand vegetable or other farm production may have returns in a matter of months (which is the shortest time horizon we are able to detect in our data, with most of the endline survey data collected multiple quarters after the cash transfers were distributed).[34]

In our view, a quick return on investment in the study setting is quite plausible given the types of investment activities that many households are engaged in. For instance, many small retail businesses purchase inventory that is sold immediately, and those who purchase livestock will soon have eggs and milk to sell. This is particularly true in this case where such investments were likely made in part to take advantage of the large spike in local spending induced by transfers to other nearby households. And even those who expand vegetable or other farm production may have returns in a matter of months (which is the shortest time horizon we are able to detect in our data, with most of the endline survey data collected multiple quarters after the cash transfers were distributed). There is no doubt that those who invest in tree crops or some other non-agricultural activities—including some training or human capital investments—might need to wait years for these investments to mature but for most of the small businesses (agricultural and non-agricultural) in the sample the time scale of returns is far more compressed.

Each of these facts is of course purely descriptive, and individually they do not fully pin down any particular theory of poverty dynamics. Collectively, though, they provide a rich set of facts for economic theory to target.

V. Extensions

A. Food security

Food security is a narrower measure of well-being than overall consumption but also of widespread humanitarian and policy interest. Recall that we prespecified as a measure of food security an index aggregating responses to questions about the number of days out of the past seven that family members experienced negative outcomes, such as skipping meals. As it is unclear whether to interpret this as a per capita or an aggregate measure, we examine results for this index alongside results for both per capita and total household food consumption. We define food consumption as the sum of expenditure on food items (including meals outside of the home) and the estimated market value of own-farm output consumed by the household.

[34] Another way to see this is to examine the relationship between deprivation (which one might expect to predict immediate consumption of transfers) and impact for households surveyed fewer vs. more months after transfer receipt. When we do this we see a similar, positive relationship between treatment effects and untreated outcomes for both groups, with the relationship if anything slightly stronger among those surveyed closer to the date of transfer receipt (Figure VI).

Regardless of which measure is used, the procedure identifies a most deprived group that is at least somewhat more deprived than the average, and than the most impacted group (Table A.8, columns (1) to (4)). In terms of per capita food consumption, for example—arguably the conceptually most appropriate measure—the most deprived group's mean consumption is 26% lower than average and 18% lower than that in the most impacted group.

The trade-off with impact is somewhat less pronounced than for financial outcomes. For the food security index itself, estimated impacts are roughly the same for the most deprived as the for the most impacted group (Table A.8, columns (7) to (9)). This is consistent with the intuitive Maslovian idea that the poorest households are both most likely to be eating too little and also most likely to spend marginal income on food. For total food consumption, however—arguably the conceptually appropriate quantity here, since households of all sizes received transfers of the same magnitude—we again see a substantial trade-off, with gains for the most impacted roughly twice as large as those for the most deprived.

Figure A.4 makes the same point visually. For the food security index (and to a lesser extent for per capita food consumption) the relationship has a negative slope, suggesting there might be little or no trade-off between targeting deprivation and impact. But when we plot effects on total food consumption against deprivation measured in per capita terms, we again see a positive relationship similar to that we observed for our financial outcomes. One might worry that this is driven by consumption of "luxury" food items such as snacks or meals out, but we obtain similar flat to upward-sloping relationships even if we restrict attention to total consumption of basic foodstuffs (e.g., staple grains).

Overall, when analysis with appropriate measures is carried out, the picture that emerges for food consumption thus seems to be that—as for financial outcomes—there is a non-trivial trade-off between targeting the most impacted and the most deprived. Because absolute impacts tend to be greater for larger households, however, this point is obscured if we only examine impacts on per capita measures of food security (including the food security index, which behaves similarly to per capita food consumption).

B. Spillover effects

An important open question of interpretation concerns the role of spillover effects. Because treatment in the experiment we study was assigned at the village level, the (differential) effects of treatment that we document on a given household *h* could in principle reflect differences in both the direct effect of transfers to household *h* itself and also indirect effects of transfers to other households in the same village. This raises some subtle issues and so, in the interests of space, we provide a full and formal exposition in the Appendix F while summarizing the main points informally here.

The key issues that arise are whether spillover effects are predictably heterogeneous with respect to the characteristics either of the households that experience them, or of the households that are treated.[35] As an example of the former, "inbound" heterogeneity, households that own businesses might benefit disproportionately when their villages are treated with cash transfers. To the extent this is because they invest their own transfers and grow their businesses, the correct inference is that reallocating transfers to them would increase average treatment effects. To the extent this is because they benefit from the shock to demand from their neighbors, however, reallocating transfers to them would have no additional effect. As an example of the latter, "outbound" heterogeneity, suppose that some households are more likely than others to employ low-income neighbors when treated; this might in principle make them attractive to target even if they would not be prioritized based on their own level of deprivation or treatment effects.

We emphasise predictability; it is quite possible, even probable, that spillover effects do vary as a function of these characteristics to some degree, but this is only relevant to the targeting problem to the extent a planner can predict the variation and hence adjust the targeting rule to reflect it. In the Appendix we conduct a series of supplemental exercises to shed some light on the ability of our predictors to capture heterogeneity in spillover effects. The first two focus on "inbound" heterogeneity: we examine whether the characteristics of ineligible households predict the effects on them of treating their village (which by definition must be spillover effects), and whether the characteristics of eligible households predict the spillover effects of treating nearby villages. The third focuses on "outbound" heterogeneity: we examine whether the average characteristics of eligible households in a village predict the effects on ineligibles of treating that village.

None of these exercises produces strong evidence of predictability. In this sense they provide some reassurance that the main results are primarily picking up heterogeneity in direct effects, and that the welfare analysis above is the appropriate one given the predictors available. At the same time, these exercises certainly do not rule out the existence of economically important heterogeneity in spillover effects, or the possibility that these might be predictable in other ways. Our general view is that reduced-form experimental identification and estimation of such effects is likely to be infeasible for the foreseeable future, mainly because the experimental designs required would be extremely expensive, could also prove politically controversial, and because it is unclear if the results produced would generalize beyond the specific context in which they are obtained.

A more plausible path forward may be to use economic reasoning to link the behavioral responses of individual households—which are relatively easy to estimate—to spillover effects. Continuing with the example above, one could measure hours of low-income labor hired in and up-weight households predicted to have large treatment effects on this outcome. This is akin to the way macroe-

[35] As we illustrate in the Appendix, any common spillover component that affects all households in a village equally would not alter our welfare analysis, since under a CARA social welfare function a common additive term does not affect the planner's ranking of treatment assignments. Under alternative social welfare functions an additional adjustment would be needed.

conomic stimulus packages in high-income countries are targeted: economists do not attempt to directly estimate how a transfer to one household will affect all other households in the economy, for example, but instead focus on estimating their marginal propensity to consume, taking this as a sufficient statistic for the impact they will have on the rest of the economy.

C. *Alternative statistical learning methods*

The analysis closes by comparing the performance of the GRF learning model that has been our focus to common alternatives. We focus on two benchmarks in particular: Ordinary Least Squares (OLS) and LASSO regression. OLS has been widely used in practice to learn scoring rules for PMT targeting, but is not designed for prediction and thus does not incorporate regularization to guard against over-fitting, which LASSO does. Neither is designed to directly learn treatment effects, as GRF does, but one can do so indirectly by predicting $\hat{Y}_h(1)$ and $\hat{Y}_h(0)$ separately and then defining $\hat{\Delta}_h = \hat{Y}_h(1) - \hat{Y}_h(0)$.[36] To give OLS and LASSO the opportunity to identify non-linearities in the data (which GRF can do without any data pre-processing) we train them on the original covariate set as well as a full set of first-order interactions.[37] The estimation procedure otherwise follows Algorithm 1 exactly.

Table 4 summarizes performance differences across methods, focusing for parsimony on properties of the groups selected as socially optimal (SO) given a moderate degree of curvature ($\alpha = 0.001$) and on consumption as the outcome.[38]

Two points are noteworthy. First, both OLS and LASSO select socially optimal groups that are somewhat less deprived than the group selected by GRF—though both predict that these groups are somewhat more deprived than they really are, OLS more so than LASSO (Panel A). This likely reflects the absence of regularization in OLS, combined with the inherent risk of a "winner's curse" in selecting extremal groups. And second, this over-optimism becomes much more pronounced when we turn to impact (Panel B). LASSO and OLS both predict that they have identified optimal groups substantially more impacted than GRF—by 61% and a whopping 169%, respectively. But in fact the average treatment effects in these groups are substantially lower than in the group selected by GRF—by over 20% in both cases. In the case of OLS, the predicted impact on the optimal group is more than double the actual.

Part of the issue may be as follows: because of the indirect way the LASSO and OLS predictions are constructed, any "noise" in the calculation of $\hat{Y}_h(0)$ and $\hat{y}_h(0)$ due for example to sampling variation will mechanically tend to generate

[36] Alternatively, one can train models on the transformed outcome $Y_h^* \equiv Y_h(T_h) - Y_h(1 - T_h)$. We obtain broadly similar results using this approach.

[37] Alternatively, one can include second order terms of continuous variables in addition to first-order interactions. OLS has worse performance with this approach while LASSO has no gains in performance.

[38] Tables C.1, C.2, and C.3 provide the full underlying estimates corresponding to the main results and for all three financial outcomes; see also Figure C.1 for a visualization of OLS and LASSO analogues to the GRF results in Figure VI.

negative correlation between $\hat{y}_h(0)$ and $\hat{\Delta}_h$, which will bias the results towards the conclusion that the most deprived are also most impacted (analogous to the problem documented by Abadie, Chingos and West (2018).) This is exactly what we see in Panel (D), where we report the overall correlations between predicted untreated values and treatment effects. This correlation is strongly positive when using GRF, but essentially zero when using LASSO and *negative* when using OLS.

One way to (loosely) summarize these results is that both LASSO and OLS—but particularly the latter, which lacks any form of regularization—think they have succeeded in identifying very highly impacted individuals, including many who are also quite deprived. As a result they predict that the planner can "have their cake and eat it too," targeting households that are very deprived while still achieving a large ATE. In reality, however, their predictions regarding impact are far too optimistic. As a result LASSO and OLS end up selecting optimal groups that are genuinely somewhat less deprived than GRF (see Panel C), but paying an unexpectedly steep cost for this in terms of actual impact. Note that given this it would not be appropriate to proceed to test hypotheses about group selection (such as those in Table 1) as these would likely over-reject.

This comparison is, of course, merely illustrative, but does suggest there is some merit in learning heterogeneous treatment effects using methods explicitly designed for that purpose. Doing otherwise may lead to two forms of error. First, policymakers may select the wrong recipients because they misjudge the trade-off between deprivation and impact. We see this here in the fact that OLS selects more of the households that are truly deprived (Panel C) but achieves a much lower average treatment effect (Panel B) than GRF. And second, conditional on the groups targeted, over-optimism about targeting performance implies over-optimism about the overall welfare gains from implementing a given targeted program. Mistakes like this will tend to distort resource allocation towards PMT-targeted programming at the expense of other approaches to targeting (or other uses of public funds entirely).

Beside these variations in the algorithm, we also consider several variations in data preparation methods as robustness checks. These address sensitivity to the discretionary choices that are needed even when (largely) using machine learning methods. We see that results are qualitatively similar if we use machine-selected covariates (via LASSO) as predictors (Tables D.1, Panel A and D.2) and if we learn deprivation using data on control eligible households only (Tables D.1, Panel B and D.3).

VI. Conclusion

This study asks whether targeting an anti-poverty program to the most "deprived" households, as is typically the case in real-world programs, has the greatest social welfare benefit, in the setting of an NGO cash transfer program in rural Kenya. A noteworthy innovation of our approach is the application of recently developed machine learning (ML) methods—specifically, generalized random forests—to learn the household characteristics that target either deprivation levels or high conditional average treatment effects across several outcomes that are prominent in international development policy debates. A central finding robust across diverse social welfare functions is that exclusively targeting the most deprived households is only attractive in a social welfare sense under very strongly redistributive preferences.

A corollary is that, for more plausible redistributive preferences, a meaningful share of the households that are social welfare maximizing to target are not those predicted to be most deprived. The results imply that policymakers should carefully consider whether automatically targeting anti-poverty assistance, like cash grants, to the poorest of the poor is necessarily appropriate in their own setting. This issue, and the results of this study, are more relevant than ever given the large rise in social assistance programming (often in the form of cash assistance) during the COVID-19 health crisis (Gentilini et al., 2020), and that in many cases appears likely to outlive the pandemic.

There are several caveats. First, the results we present apply to large-scale cash grants, but patterns of impact, and the nature of the deprivation-impact tradeoff, may plausibly differ for other types of assistance (e.g., subsidized credit or public health insurance). The rural Kenyan setting we study is also ethnically and religiously homogeneous and characterized by relatively limited inequality across households (within a village); for instance, the vast majority of households are landowners. In other settings with greater gaps in household wealth and living standards, or more salient social divisions, the benefits to targeting the poorest may be more pronounced. At the same time, in such settings, the gains from targeting those with the largest treatment effects may also be greater, and it is unclear which of these two effects outweighs the other.

Second, we measure endline outcomes (and thus treatment effects) over a substantial but still limited time window. Our data coverage begins shortly after transfer receipt and continues for nearly two years, which we see as a meaningful advance relative to past work on targeting, the bulk of which has had to limit itself to data collected at a single point in time. But both targeting performance and the persistence of cash impacts might of course change over yet longer time horizons (Kondylis and Loeser, 2021). The longer-term effects of this particular cash transfer program are the subject of ongoing work (Egger et al., 2021).

Third, we caution that targeting assistance to those with the largest treatment effects may deepen existing inequalities. It appears that several marginalized subgroups in the population we study, e.g., widow-headed households or those with few or no prime-age adults, translate the cash grants into less substantial gains in future consumption, assets and income. It is possible that this finding might hold more generally: groups that are frequently marginalized or discriminated against (e.g., women, and ethnic or religious minorities, etc.) may not be able to leverage an assistance program as effectively as more favored groups that have other social advantages. The analytical approach we propose might, in this case, conclude that it is social welfare optimal to target assistance to precisely these favored groups, even though this decision to target assistance to those who would use it "effectively" will tend to reinforce existing social inequalities. One strategy to address this, which we illustrated above, is to incorporate Pareto weights for marginalized groups into the social welfare function. But a potentially more cost-effective approach is to design alternative programs that generate larger benefits for these groups. Sustained assistance over a longer period of time or at higher levels, or in a different form, might be needed to allow deprived and marginalized groups to take full advantage of the opportunities provided by an aid program. This is beyond the scope of our study given the one-time transfer and the limited time frame we examine, but could be a rationale for more aggressively targeting assistance to deprived groups, providing complementary forms of assistance, or extending cash assistance over longer time periods (as in an ongoing universal basic income study in the same region, Banerjee et al., 2020). The correct inference, in other words, might be akin to the analogous idea (in the separate microcredit literature) in Morduch (1999) that "poorer households should be served by other interventions than credit" if they benefit less from credit, rather than writing them off entirely.

Despite these important limitations, we hope the approach proposed in this study can be used to reinvigorate real-world policy discussions around optimal targeting of social assistance. The use of richer data and sophisticated machine learning methods to target the households that are most likely to contribute to social welfare could potentially even help to build greater popular support for anti-poverty programs by convincing the electorate that social benefits are being maximized (rather than targeting being driven by politicians' electoral considerations, say), although it may be a challenge to transparently and succinctly explain ML methods to many citizens. Doing so effectively might even make such programs more politically sustainable. In our view, it will be valuable to extend the approach in this study to other forms of assistance (beyond cash transfers), to other contexts, and to the use of alternative machine learning methods, and to ensure an active feedback loop with international development policymakers. But at a minimum, we hope the results of this study lead real-world policymakers to more systematically gather evidence on program impacts, and to consider that continuing to reflexively target aid programs solely to the most deprived may not always maximize social welfare.

REFERENCES

Abadie, Alberto, Matthew M. Chingos, and Martin R. West. 2018. "Endogenous Stratification in Randomized Experiments." *The Review of Economics and Statistics*, 100(4): 567–580.

Alatas, Vivi, Abhijit Banerjee, Rema Hanna, Benjamin A. Olken, and Julia Tobias. 2012. "Targeting the Poor: Evidence from a Field Experiment in Indonesia." *American Economic Review*, 102(4): 1206–40.

Alesina, Alberto, and Dani Rodrik. 1994. "Distributive Politics and Economic Growth." *Quarterly Journal of Economics*, 109(2): 465–490.

Ambuehl, Sandro, B. Douglas Bernheim, and Axel Ockenfels. 2021. "What Motivates Paternalism? An Experimental Study." *American Economic Review*, 111(3): 787–830.

Andrews, Isaiah, Toru Kitagawa, and Adam McCloskey. 2021. "Inference on Winners." Working paper.

Athey, Susan, and Stefan Wager. 2021. "Policy Learning With Observational Data." *Econometrica*, 89(1): 133–161.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized random forests." *The Annals of Statistics*, 47(2): 1148 – 1178.

Babcock, Bruce A, E Kwan Choi, and Eli Feinerman. 1993. "Risk and probability premiums for CARA utility functions." *Journal of Agricultural and Resource Economics*, 17–24.

Baird, Sarah, Craig McIntosh, and Berk Özler. 2011. " Cash or Condition? Evidence from a Cash Transfer Experiment." *The Quarterly Journal of Economics*, 126(4): 1709–1753.

Balakrishnan, Uttara, Johannes Haushofer, and Pamela Jakiela. 2020. "How soon is now? Evidence of present bias from convex time budget experiments." *Experimental Economics*, 23(2): 294–321.

Banerjee, Abhijit, Dean Karlan, and Jonathan Zinman. 2015. "Six Randomized Evaluations of Microcredit: Introduction and Further Steps." *American Economic Journal: Applied Economics*, 7(1): 1–21.

Banerjee, Abhijit, Michael Faye, Alan Krueger, Paul Niehaus, and Tavneet Suri. 2020. "Effects of a Universal Basic Income during the pandemic." Working Paper.

Banerjee, Abhijit V., Paul J. Gertler, and Maitreesh Ghatak. 2002. "Empowerment and Efficiency: Tenancy Reform in West Bengal." *Journal of Political Economy*, 110(2): 239–280.

Barseghyan, Levon, Francesca Molinari, Ted O'Donoghue, and Joshua C Teitelbaum. 2018. "Estimating risk preferences in the field." *Journal of Economic Literature*, 56(2): 501–64.

Bastagli, Francesca, Jessica Hagen-Zanker, Luke Harman, Valentina Barca, Georgina Sturge, Tanja Schmidt, and Luca Pellerano. 2016. "Cash transfers: what does the evidence say? A rigorous review of programme impact and of the role of design and implementation features."

Basurto, Maria Pia, Pascaline Dupas, and Jonathan Robinson. 2020. “Decentralization and efficiency of subsidy targeting: Evidence from chiefs in rural Malawi.” *Journal of Public Economics*, 185(104047).

Bertrand, Marianne, Bruno Crépon, Alicia Marguerie, and Patrick Premand. 2021. “Do Workfare Programs Live Up to Their Promises? Experimental Evidence from Cote D’Ivoire.” NBER Working Paper No. 28664.

Bhattacharya, Debopam, and Pascaline Dupas. 2012. “Inferring welfare maximizing treatment assignment under budget constraints.” *Journal of Econometrics*, 167(1): 168–196.

Björkegren, Daniel, Joshua E. Blumenstock, and Samsun Knight. 2022. “(Machine) Learning What Policies Value.” arXiv.

Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. 2015. “Predicting poverty and wealth from mobile phone metadata.” *Science*, 350(6264): 1073–1076.

Brown, Caitlin, Martin Ravallion, and Dominique van de Walle. 2018. “A poor means test? Econometric targeting in Africa.” *Journal of Development Economics*, 134: 109–124.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. 2018. “Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India.” NBER Working Paper No. 24678.

Chetty, Raj. 2006. “A new method of estimating risk aversion.” *American Economic Review*, 96(5): 1821–1834.

Deaton, Angus, and Christina Paxson. 1998. “Economies of Scale, Household Size, and the Demand for Food.” *Journal of Political Economy*, 106(5): 897–930.

de Mel, Suresh, David McKenzie, and Christopher Woodruff. 2008. “Returns to Capital in Microenterprises: Evidence from a Field Experiment.” *The Quarterly Journal of Economics*, 123(4): 1329–1372.

Demirgüç-Kunt, Asli, Leora Klapper, Dorothe Singer, and Saniya Ansar. 2022. *The Global Findex Database 2021: Financial inclusion, digital payments, and resilience in the age of COVID-19.* World Bank Publications.

DiCiccio, Thomas J, and Bradley Efron. 1996. “Bootstrap confidence intervals.” *Statistical science*, 11(3): 189–228.

Diciccio, Thomas J, and Joseph P Romano. 1988. “A review of bootstrap confidence intervals.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 50(3): 338–354.

Egger, Dennis, Johannes Haushofer, Edward Miguel, and Michael Walker. 2021. "GE Effects of Cash Transfers: Pre-analysis plan for Endline 2 Household Welfare Analyses." AEA Trial Registry.

Egger, Dennis, Johannes Haushofer, Edward Miguel, Paul Niehaus, and Michael Walker. 2022*a*. "Data and Code for: "General Equilibrium Effects of Cash Transfers: Experimental Evidence from Kenya"'." Available at: `https://onlinelibrary.wiley.com/doi/full/10.3982/ECTA17945`.

Egger, Dennis, Johannes Haushofer, Edward Miguel, Paul Niehaus, and Michael Walker. 2022*b*. "General Equilibrium Effects of Cash Transfers: Experimental Evidence from Kenya." *Econometrica*, 90: 2603 – 2643.

Egger, Dennis, Johannes Haushofer, Edward Miguel, Paul Niehaus, and Michael Walker. 2024. "Trial Registry: General Equilibrium Effects of Cash Transfers in Kenya."

Elminejad, Ali, Tomas Havranek, and Zuzana Irsova. 2022. "Relative risk aversion: a meta-analysis."

Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *The Quarterly Journal of Economics*, 114(3): 817–868.

Gentilini, Ugo, Mohamed Almenfi, Ian Orton, and Pamela Dale. 2020. "Social protection and jobs responses to COVID-19." World Bank, Washington, DC.

Hanna, Rema, and Benjamin A. Olken. 2018. "Universal Basic Incomes versus Targeted Transfers: Anti-Poverty Programs in Developing Countries." *Journal of Economic Perspectives*, 32(4): 201–26.

Haushofer, Johannes, and Jeremy Shapiro. 2016. "The Short-term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya." *The Quarterly Journal of Economics*, 131(4): 1973–2042.

Hussam, Reshmaan, Natalia Rigol, and Benjamin N Roth. 2022. "Targeting High Ability Entrepreneurs Using Community Information: Mechanism Design In The Field." *American Economic Review*, 112: 861–898.

Kitagawa, Toru, and Aleksey Tetenov. 2018. "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice." *Econometrica*, 86(2): 591–616.

Kondylis, Florence, and John Loeser. 2021. "Intervention Size and Persistence." World Bank Policy Research WPS 9769.

Lindbeck, Assar, and Jörgen W. Weibull. 1987. "Balanced-budget redistribution as the outcome of political competition." *Public Choice*, 52(3): 273–297.

Manacorda, Marco, Edward Miguel, and Andrea Vigorito. 2011. "Government Transfers and Political Support." *American Economic Journal: Applied Economics*, 3(3): 1–28.

Manski, Charles F. 2004. "Statistical Treatment Rules for Heterogeneous Populations." *Econometrica*, 72(4): 1221–1246.

Maslow, A H. 1943. "A theory of human motivation." *Psychological Review*, 50(4): 370–396.

McKenzie, David, and Dario Sansone. 2019. "Predicting entrepreneurial success is hard: Evidence from a business plan competition in Nigeria." *Journal of Development Economics*, 141: 102369.

Meager, Rachael. 2022. "Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature." *American Economic Review*, 112: 1818–47.

Morduch, Jonathan. 1999. "The Microfinance Promise." *Journal of Economic Literature*, 37(4): 1569–1614.

Nelson, Julie A. 1988. "Household Economies of Scale in Consumption: Theory and Evidence." *Econometrica*, 56(6): 1301–1314.

Persson, Torsten, and Guido Tabellini. 1994. "Is Inequality Harmful for Growth?" *American Economic Review*, 84(3): 600–621.

Premand, Patrick, and Pascale Schnitzer. 2021. "Efficiency, Legitimacy, and Impacts of Targeting Methods: Evidence from an Experiment in Niger." *World Bank Economic Review*, 35(4): 892–920.

Rabin, Matthew. 2000. "Risk Aversion and Expected-Utility Theory: A Calibration Theorem." *Econometrica*, 68(5): 1281–1292.

Saez, Emmanuel, and Stefanie Stantcheva. 2016. "Generalized Social Marginal Welfare Weights for Optimal Tax Theory." *American Economic Review*, 106(1): 24–45.

Sen, Amartya. 1999. *Development as freedom.* Oxford:OUP.

Subramanian, Shankar, and Angus Deaton. 1996. "The Demand for Food and Calories." *Journal of Political Economy*, 104(1): 133–162.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society, Series B (Methodological)*, 58(1): 267–288.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association*, 113(523): 1228–1242.

Walker, Michael. 2018. “Informal taxation and cash transfers: Experimental evidence from Kenya.” Working paper.

World Bank. 2022. “Fact sheet: an adjustment to global poverty lines.”

World Bank. n.d.. “PPP conversion factors for GDP (LCU per international $).” Available at: `<https://data.worldbank.org/indicator/PA.NUS.PPP>`.Accessed August 2, 2023.

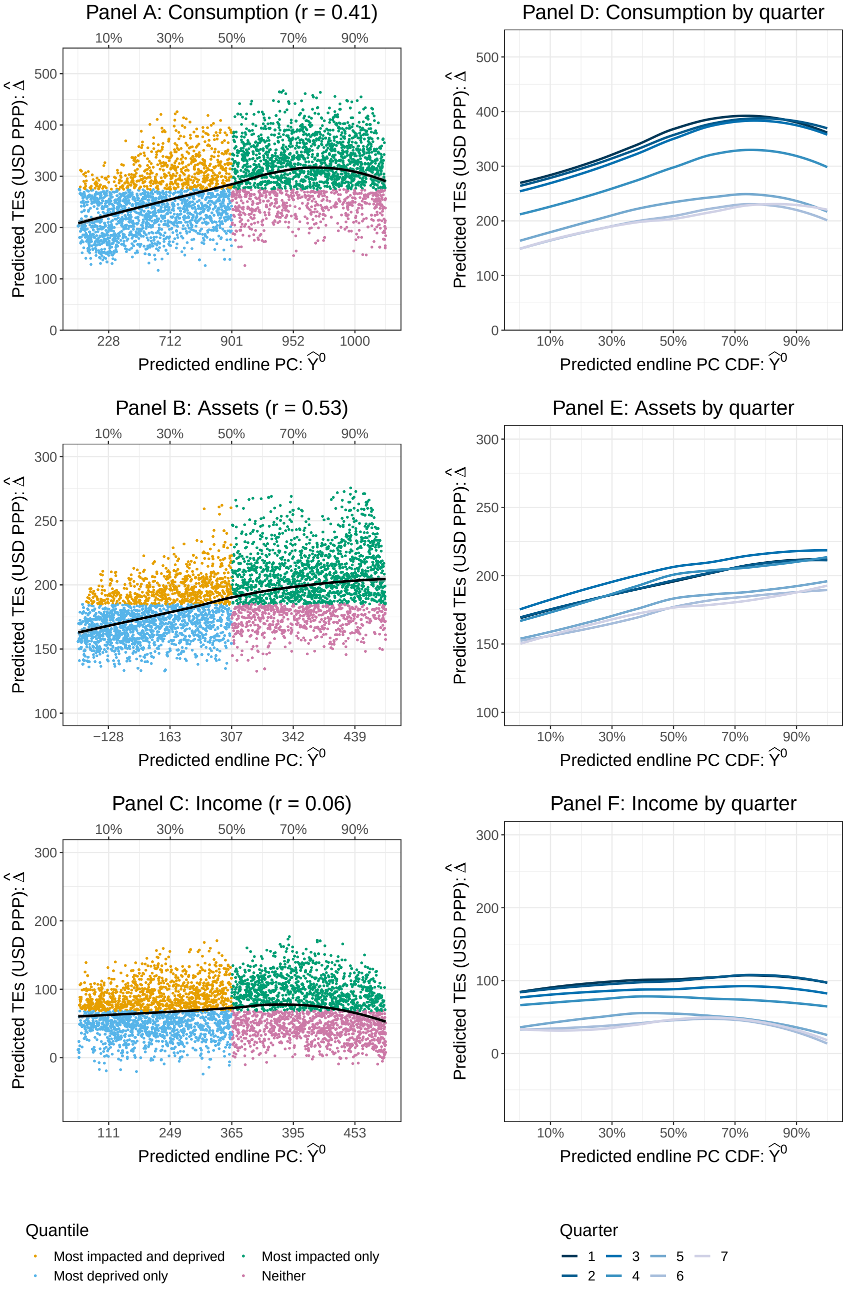

Figure 1. Predicted treatment effects ($\hat{\Delta}_h$) plotted against the predicted untreated per capita values ($\hat{y}_h^0$)

Notes: Each sub-figure plots predicted treatment effects for an outcome (y-axis) against the predicted endline values (x-axis). Panels A, B and C include scatter plots of the household-level estimates and a local regression line, and are color-coded according to their deprivation and impact classification status. The correlation (r) between predicted endline values and treatment effects for the median model is reported in the subfigure title. Panels D, E, and F plot the local regression lines generated from data for each quarter after treatment. As we generate 150 models per outcome, the figures presented are from the median model in terms of the difference in average treatment effects between the most deprived and most impacted groups for each outcome. Both predicted endline and predicted treatment effects are estimated from generalized random forest models with the same set of covariates. Predicted endline values and treatment effects are from models trained on time-demeaned data; a constant was added to the predicted endline outcomes so that the overall predicted mean matches the observed sample mean. Monetary values are in USD PPP (2016).

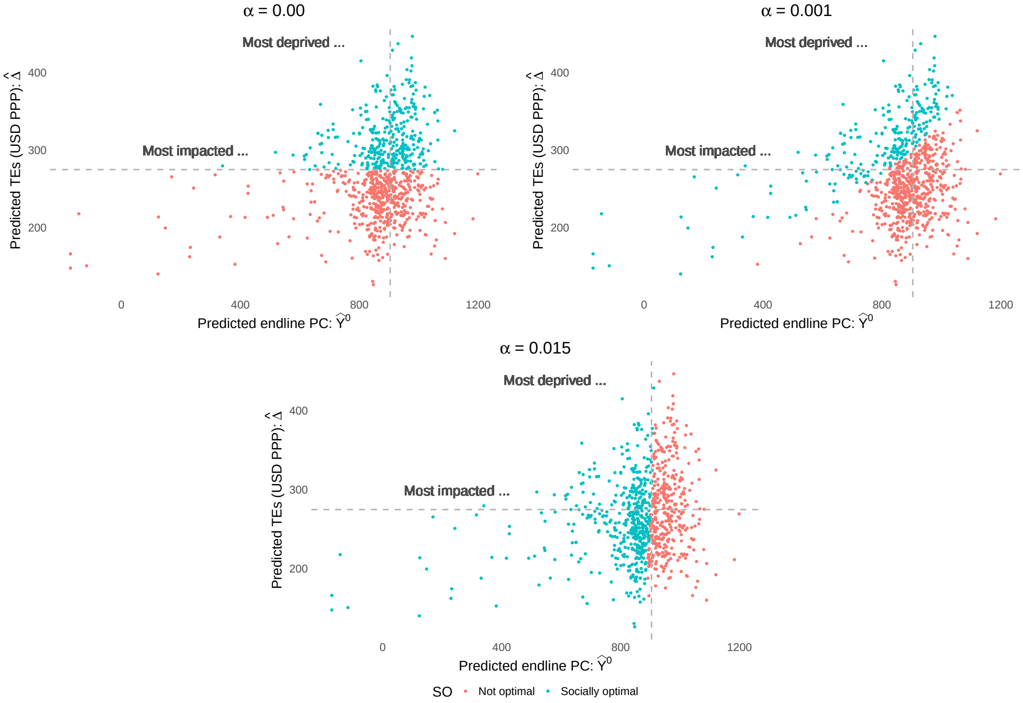

FIGURE 2. PLOTTING PREDICTED DEPRIVATION VERSUS IMPACT BY SOCIALLY OPTIMAL STATUS

Notes: This figure plots the predicted endlines and predicted treatment effects for households of 4 members (the median size). Socially optimal groups are highlighted for different curvature values using CARA. Dashed lines denote the thresholds for the most impacted and most deprived households. For exposition purposes, socially optimal households were selected without cross-fitted thresholds, using integrated predictions across quarters (static models), for households of the same size. A constant was added to the predicted endline outcomes so that the overall predicted mean matches the observed sample mean, since GRF models were trained with time-demeaned data. Monetary values are in USD PPP (2016).

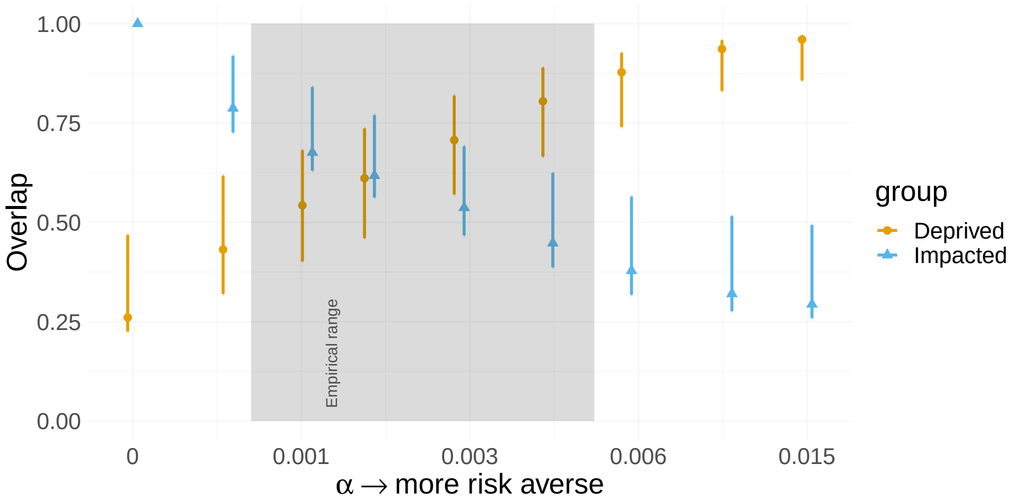

Figure 3. Overlap of socially optimal households to target with most deprived and most impacted

Notes: The figure plots point estimates and bootstrapped 95% confidence intervals of the share of I (D) households that are also "socially optimal" for a planner to treat. Socially optimal households are those in the top 50% of households ranked by potential gains from treatment using a CARA utility function. The shaded area denotes the range of values of the absolute risk aversion parameter α that imply a relative risk aversion (CRRA) parameter ρ in the range $[0.5, 4]$ evaluated at mean consumption per capita in our sample ($\rho = \alpha c$, where c is consumption). In Elminejad, Havranek and Irsova (2022) the authors conduct a meta-analysis of 92 studies and find a mean estimate of $\rho = 1$ in economics articles, with most of the estimate mass lying in the $[0, 4]$ range. Moreover, Chetty (2006) finds that only values of $\rho < 2$ rationalize established facts about the labor market in a wide range of contexts, and finds a mean estimate of $\rho = 1$ across 13 studies using data from various countries and samples. The x-axis is transformed by an inverse arc-sine function to emphasize the range of values that are closer to the empirical estimates in the literature.

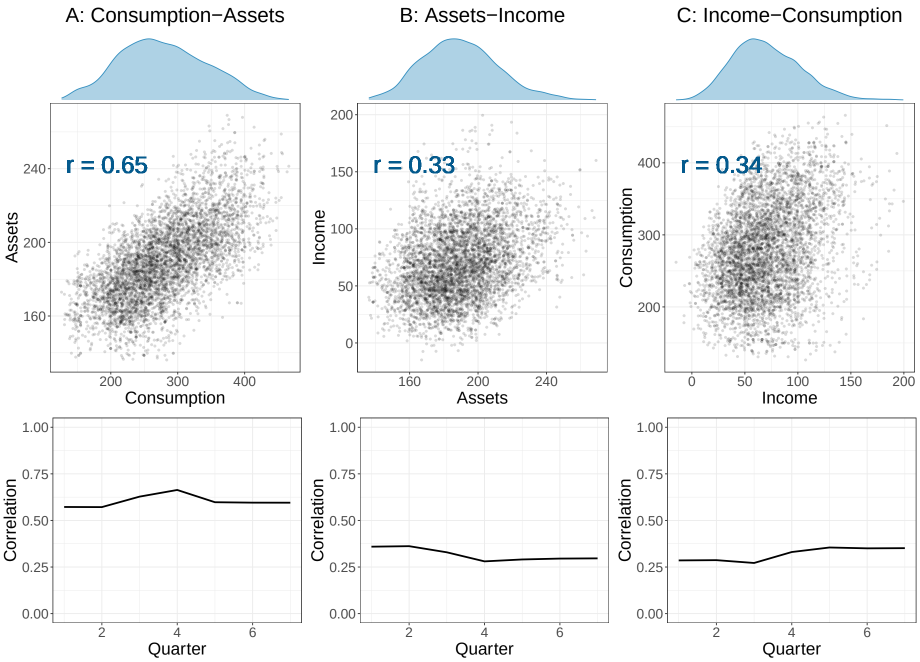

Figure 4. Cross-outcome relationships in predicted treatment effects

Notes: This figure looks at correlations in predicted treatment effects across different outcomes for the main models presented in Figure VI and Table 2. Panel A (column 1) looks at the relationship between consumption and assets, Panel B (column 2) looks at the relationship between assets and income, and Panel C (column 3) looks at income and consumption, with the former variable plotted along the x-axis. Monetary values are in USD PPP (2016). The top row plots the kernel densities of treatment effects for the x-axis, while the middle row shows scatter plots of predicted treatment effects. The bottom row plots the pairwise correlation between the variables by quarter since treatment. For each household we use the average prediction across the 150 models trained. r denotes the correlation between the predicted treatment effects for the median model.

Table 1—Main social welfare analysis: overlap of socially optimal households to target with most deprived and most impacted

	(1)	(2)	(3)	(4)	(5)
		Most deprived		Most impacted	
	CE	Share	p-val	Share	p-val
			$D>0.95$		$I>0.95$
Panel A: Consumption, CARA					
$\alpha = 0.0000$	\$50	0.26	0.00	1.00	1.00
$\alpha = 0.0005$	\$49	0.43	0.00	0.79	0.00
$\alpha = 0.0010$	\$49	0.54	0.00	0.68	0.00
$\alpha = 0.0075$	\$41	0.91	0.01	0.35	0.00
$\alpha = 0.0150$	\$33	0.96	0.31	0.29	0.00
Panel B: Other welfare measures, CARA					
Assets, $\alpha = 0.0010$	\$49	0.63	0.00	0.61	0.00
Income, $\alpha = 0.0010$	\$49	0.54	0.00	0.89	0.23
Panel C: Sensitivity checks on consumption					
CRRA, $\rho = 0.5$		0.36	0.00	0.87	0.04
CRRA, $\rho = 2$		0.53	0.00	0.68	0.00
Time discounting: $\beta = 15\%$, $\alpha = 0.0001$		0.54	0.00	0.67	0.00
Re-targeting dynamics, $\alpha = 0.001$		0.53	0.00	0.68	0.00
Pareto weights, $\alpha = 0.0005$		0.52	0.00	0.67	0.00
Saez-Stantcheva (2016), $\alpha = 0.0005$		0.52	0.00	0.67	0.00

Notes: Column 1 denotes the certainty equivalent (CE) of a 50-50 lottery over \$0 or \$100 under the specified CARA α parameter value. Column 2 (4) reports the share of households belonging to D (I) that are also "socially optimal" (those in the top 50% of households ranked by potential gains from treatment) for a planner to treat for a given utility function (CARA or CRRA) and parameter value (α or ρ). Reported shares are the mean of 150 5-fold GRF iterations; median ratios are similar (not shown). Columns (3) and (5) report p-values testing whether a planner would prefer to predominately target only the most deprived (D) or the most impacted (I). Panel C presents a variety of sensitivity analyses. For additional sensitivity checks, parameter values, and outcomes see Appendix tables A.1 (assets and income), A.2 (CRRA), A.3 (observable assets), C.1 (OLS and LASSO based prediction models), D.1 (additional robustness checks), and G.2 (pareto weights).

Table 2—Characterizing the socially optimal (SO), most deprived (D) and most impacted (I) groups

	(1)	(2)	(3)	(4)	(5)	(6)
Statistic	All	(D)	(SO)	(I)	(D)-(I)	Inference
Panel A1: Consumption, predicted per capita untreated outcomes (y_h^0)						
Predicted	750	542	619	923	-381	
Actual	729	503	554	911	-408	(-466,-197) [-469,-344] p < 0.01
Panel A2: Consumption, average treatment effects (Δ_h)						
Predicted	277	247	303	326	-79	
Actual	310	247	439	405	-159	(-349,-46) [-321,-3] p: 0.01
Panel B1: Assets, predicted per capita untreated outcomes (y_h^0)						
Predicted	232	85	128	343	-258	
Actual	213	53	96	336	-283	(-308,-96) [-313,-253] p < 0.01
Panel B2: Assets, average treatment effects (Δ_h)						
Predicted	189	178	195	207	-29	
Actual	182	154	167	188	-34	(-168,-6) [-123,58] p: 0.03
Panel C1: Income, predicted per capita untreated outcomes (y_h^0)						
Predicted	304	186	258	321	-135	
Actual	297	170	247	323	-153	(-229,-65) [-182,-127] p < 0.01
Panel C2: Income, average treatment effects (Δ_h)						
Predicted	69	66	92	94	-28	
Actual	85	79	107	94	-15	(-224,5) [-131,91] p: 0.07

Notes: This table presents summary statistics of inputs into the welfare analysis, namely the group averages of actual and predicted per capita endline values among transfer-eligible households in treatment and control villages (y_h^0, panels A1, B1, C1) and treatment effects for transfer-eligible households (Δ_h, panels A2, B2, C2). The socially optimal (SO) group is calculated with a value of $\alpha = 0.001$. We report the 95% BCa CI for the actual difference statistic computed through empirical bootstrap for the whole procedure in parentheses, and the BCa CI computed through empirical bootstrap conditional on the GRF model predictions in brackets. We also report the *p*-values corresponding to the standard errors in parenthesis (bootstrap on the whole procedure). $N = 2,367$ for y_h^0 and $N = 4,749$ for Δ_h. Monetary values are in USD PPP (2016).

Table 3—Variable importance for predicting (non-)deprivation and impacts

		Predicted untreated outcomes (y_h^0)			Predicted treatment effects (Δ_h)		
Variable	Mean	Consumption	Assets	Income	Consumption	Assets	Income
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Household demographics							
HH size	4.38	**0.68 (1,+)**	**0.71 (1,+)**	**0.40 (1,+)**	**0.18 (1,+)**	**0.19 (1,+)**	**0.17 (1,+)**
Female head	0.69	0.01 (11,-)	0.00 (13,-)	0.03 (5,-)	0.05 (6,+)	0.05 (5,+)	0.05 (6,-)
Has children	0.81	**0.06 (3,+)**	**0.05 (3,+)**	0.05 (4,+)	0.01 (15,-)	0.01 (16,-)	0.01 (14,-)
Has children in school	0.66	0.02 (5,+)	0.01 (7,+)	0.01 (9,+)	0.03 (10,+)	0.03 (11,+)	0.03 (10,-)
Has child under 3	0.50	0.00 (16,+)	0.00 (15,-)	0.00 (15,+)	0.04 (8,+)	0.05 (7,+)	0.05 (7,+)
Has child under 6	0.64	0.00 (15,+)	0.01 (8,+)	0.00 (12,+)	0.03 (13,+)	0.03 (10,+)	0.03 (13,-)
Widow	0.21	0.05 (4,-)	0.02 (5,-)	**0.20 (3,-)**	0.01 (14,+)	0.02 (12,-)	0.01 (15,+)
Has elder member	0.11	**0.09 (2,-)**	0.02 (4,-)	**0.22 (2,-)**	0.01 (16,+)	0.01 (15,+)	0.00 (16,-)
Panel B: Financial characteristics							
Employed	0.34	0.00 (14,-)	0.01 (11,-)	0.00 (11,+)	0.03 (9,+)	0.04 (8,+)	0.06 (4,+)
Self-employed	0.27	0.01 (10,+)	0.01 (10,+)	0.03 (6,+)	0.05 (5,-)	0.05 (6,+)	**0.06 (3,-)**
Has any livestock	0.26	0.00 (12,+)	**0.10 (2,+)**	0.00 (13,+)	**0.07 (3,+)**	**0.09 (2,+)**	0.05 (8,+)
Owns land	0.84	0.01 (9,-)	0.00 (14,-)	0.00 (14,-)	0.03 (12,+)	0.02 (13,+)	0.03 (12,+)
Owns 1/4 acre	0.82	0.00 (13,-)	0.00 (16,-)	0.00 (16,+)	0.03 (11,+)	0.02 (14,+)	0.03 (11,+)
Owns TV or radio	0.62	0.01 (6,+)	0.02 (6,+)	0.01 (10,+)	0.04 (7,-)	0.04 (9,+)	0.04 (9,-)
Meals yesterday	2.29	0.01 (7,+)	0.01 (9,+)	0.01 (8,+)	**0.07 (2,-)**	**0.08 (3,-)**	**0.07 (2,+)**
Meals with protein yesterday	0.43	0.01 (8,+)	0.00 (12,+)	0.01 (7,+)	0.05 (4,-)	0.06 (4,+)	0.06 (5,+)
Panel C: Study variables							
Treatment	0.50	0.01 (+)	0.01 (+)	0.00 (+)			
Months since treated	19.09	0.03 (-)	0.02 (+)	0.02 (-)	0.28 (-)	0.21 (-)	0.25 (-)

Notes: Column (1) reports the unconditional mean of each variable at the baseline. Columns (2)-(5) report variable importance for endline predictions, and columns (6)-(9) report importance for predicted treatment effects. Variable importance is measured as the a depth-weighted average of the share of splits created in the process of growing trees that split on a particular variable (see Equation (15)). The first argument in parentheses is the variable importance ranking; the second argument is whether the predicted outcome increases (+) or decreases (−) when the variable is 1 versus 0 for indicators or a one standard deviation increase from the mean for continuous variables, fixing all other covariates to their mean. For each outcome, the top three variables by importance are in bold. $N = 4,749$.

Table 4—Comparison across prediction methods for consumption

Statistic:	GRF (1)	LASSO (2)	OLS (3)
Panel A: Untreated outcome (per capita)			
Predicted y_h^0 for (SO)	619	644	626
Actual y_h^0 for (SO)	554	658	700
5% and 95% quantiles of predicted y_h^0	(-171, 1012)	(-204,1099)	(-430, 1246)
Panel B: Treatment effect			
Predicted Δ_h for (SO)	303	487	815
Actual Δ_h for (SO)	439	324	346
5% and 95% quantiles of predicted Δ_h	(196, 374)	(-104,712)	(-766, 1389)
Panel C: Comparison to observed untreated outcome (per capita)			
Proportion actual (D) in selected (D)	0.633	0.631	0.608
Proportion actual (D) in selected (SO)	0.561	0.536	0.507
Panel D: correlation between the untreated outcome and the treatment effect			
$\rho(y_h^0, \Delta_h)$	0.41	0.02	-0.18

Notes: This table presents comparisons across methods for learning predictions using consumption as our outcome of interest. Column (1) presents our main estimates using generalized random forests (GRF), as in Table 2. Columns (2) and (3) show results using LASSO (as in Table C.3) and OLS (as in Table C.2), respectively. Panel A presents results by group for the untreated outcome (per capita), while Panel B presents treatment effects by group. Panels A and B report the predicted and actual group means for the socially optimal (SO) group, as defined in the welfare results using CARA utility for $\alpha = 0.001$ (as in Table 1 for Column 1). The socially optimal group thus varies based on the predictions generated by each method. Moreover, panels A and B report the 5% to 95% quantile range of the predictions for y_h^0 and Δ_h. Panel C uses endline survey data from control group villages to define the group that is observed to be the most deprived as households with per-capita consumption below the median, and compares this group (the "actual D" group) to their assignments under different learning methods. For more details on each of these methods and results for other outcomes for Panels A and B, see the tables referenced above. Panel d shows the correlations between untreated per capita outcomes and treatment effects. Monetary values for panels A and B are in USD PPP (2016).