

The Millennium Villages Project: a retrospective, observational, endline evaluation

Shira Mitchell, Andrew Gelman, Rebecca Ross, Joyce Chen, Sehrish Bari, Uyen Kim Huynh, Matthew W Harris, Sonia Ehrlich Sachs, Elizabeth A Stuart, Avi Feller, Susanna Makela, Alan M Zaslavsky, Lucy McClellan, Seth Ohemeng-Dapaah, Patricia Namakula, Cheryl A Palm, Jeffrey D Sachs



Summary

Background The Millennium Villages Project (MVP) was a 10 year, multisector, rural development project, initiated in 2005, operating across ten sites in ten sub-Saharan African countries to achieve the Millennium Development Goals (MDGs). In this study, we aimed to estimate the project's impact, target attainment, and on-site spending.

Methods In this endline evaluation of the MVP, we retrospectively selected comparison villages that best matched the project villages on possible confounding variables. Cross-sectional survey data on 40 outcomes of interest were collected from both the project and the comparison villages in 2015. Using these data, as well as on-site spending data collected during the project, we estimated project impacts as differences in outcomes between the project and comparison villages; target attainment as differences between project outcomes and prespecified targets; and on-site spending as expenditures reported by communities, donors, governments, and the project. Spending data were not collected in the comparison villages.

Findings Averaged across the ten project sites, we found that impact estimates for 30 of 40 outcomes were significant (95% uncertainty intervals [UIs] for these outcomes excluded zero) and favoured the project villages. In particular, substantial effects were seen in agriculture and health, in which some outcomes were roughly one SD better in the project villages than in the comparison villages. The project was estimated to have no significant impact on the consumption-based measures of poverty, but a significant favourable impact on an index of asset ownership. Impacts on nutrition and education outcomes were often inconclusive (95% UIs included zero). Averaging across outcomes within categories, the project had significant favourable impacts on agriculture, nutrition, education, child health, maternal health, HIV and malaria, and water and sanitation. A third of the targets were met in the project sites. Total on-site spending decreased from US\$132 per person in the first half of the project (of which \$66 was from the MVP) to \$109 per person in the second half of the project (of which \$25 was from the MVP).

Interpretation The MVP had favourable impacts on outcomes in all MDG areas, consistent with an integrated rural development approach. The greatest effects were in agriculture and health, suggesting support for the project's emphasis on agriculture and health systems strengthening. The project conclusively met one third of its targets.

Funding The Open Society Foundations, the Islamic Development Bank, and the governments of Japan, South Korea, Mali, Senegal, and Uganda.

Copyright © The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND 4.0 license.

Introduction

In September, 2000, world leaders at the UN Millennium Summit committed their nations to reducing extreme poverty and set targets with a deadline of 2015, known as the Millennium Development Goals (MDGs).¹ The UN Millennium Project, initiated by then UN Secretary-General Kofi Annan and directed by Jeffrey Sachs, estimated that interventions to achieve the MDGs would require a 10 year average investment of between \$101 and \$127 per person per year (all monetary units given as 2005 US\$ per person per year), with low-income settings requiring \$58–80 from external financing.^{2–5} Motivated by the UN Millennium Project's findings, the Millennium Villages Project (MVP) was initiated in 2005 to achieve the MDGs within 5 years in rural villages in sub-Saharan Africa, focusing on quick-win interventions.⁶ Thereafter, additional funding enabled the MVP to be redesigned as a

10 year project, including longer-term strategies to achieve the MDGs by 2015.⁷ The project expanded to 14 sites across ten countries by the end of 2006. Countries were selected on the basis of political stability and government commitment to the MDGs. Millennium Village (MV) sites were selected from rural areas of high under-nutrition, representing varied agroecological zones, and with local political buy-in and community ownership.^{8,9} The project implemented integrated interventions in poverty, agriculture, nutrition, education, health, and infrastructure (panel; appendix), and aimed to contribute \$60 (of \$80) in external financing, consistent with the UN Millennium Project's recommendations.

The project expanded its area of coverage in ten of the 14 MV sites (one in each country; figure 1) to include at least 25 000 inhabitants per site, enabling reduced management costs per capita. Four sites were not scaled

Lancet Glob Health 2018;
6: e500–13

See [Comment](#) pages e470 and e472

Mathematica Policy Research, Princeton, NJ, USA (S Mitchell PhD); Statistics Department (Prof A Gelman PhD, S Makela PhD), and Center for Sustainable Development (J Chen MSSW, S E Sachs MD), Earth Institute (R Ross MSPH, S Bari MPH, M W Harris MA, L McClellan MIA, Prof J D Sachs PhD), Columbia University, New York, NY, USA; Millennium Promise, New York, NY, USA (U K Huynh PhD); Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA (Prof E A Stuart PhD); Goldman School of Public Policy, UC Berkeley, Berkeley, CA, USA (A Feller PhD); Health Care Policy Department, Harvard Medical School, Boston, MA, USA (Prof A M Zaslavsky PhD); Millennium Promise, Dakar, Senegal

(S Ohemeng-Dapaah PhD); Millennium Promise, Kampala, Uganda (P Namakula MSc); and Institute for Sustainable Food Systems, Department of Agricultural and Biological Engineering, University of Florida, Gainesville, FL, USA (Prof C A Palm PhD)

Correspondence to:
Dr Shira Mitchell, Mathematica Policy Research, Princeton, NJ 08543, USA
sam942@mail.harvard.edu

See [Online](#) for appendix

Research in context

Evidence before this study

Since its launch in 2005, the Millennium Villages Project (MVP) has been both admired and scrutinised for its implementation and evaluation. Its ambitious objective to achieve the Millennium Development Goals (MDGs) in rural African villages (within an initial timeframe of only 5 years) attracted substantial attention. Previous mid-term evaluations of the project insufficiently addressed methodological challenges of impact estimation, particularly the choice of a comparison group.

Added value of this study

In this endline evaluation of the MVP, we aimed to estimate project impact, target attainment, and on-site spending. To estimate impact, we used a rigorous and pre-registered procedure to retrospectively select comparison villages that best matched the project villages on possible confounding variables. We used both classical and Bayesian methods to

synthesise information from 40 MDG-related outcomes. Our evaluation is the first to assess the project after the full 10 years of implementation. This study exemplifies methods for retrospective observational studies, which enable learning about the effects of policies and projects implemented without a prospectively designed evaluation and without random assignment to treatment.

Implications of all available evidence

Although the project's goal to achieve all of the MDGs was not met, we estimated that, averaged across the ten sites, the MVP had a significant effect on 30 of 40 outcomes of interest, all of which favoured the project villages. The greatest effects were on agriculture and health outcomes, supporting the project's approach of agriculture and health systems strengthening, and less conclusive for impacts on poverty, nutrition, and education outcomes.

Panel: Examples of Millennium Villages Project interventions

Poverty

- Enterprise diversification
- Business development

Agriculture

- Promotion and subsidisation of improved fertilisers and seeds
- Agronomy training
- Installation of irrigation systems

Nutrition

- Micronutrient supplementation
- Management of acute malnutrition

Education

- Construction of schools and teacher housing
- Teacher training
- Provision of school meals
- Upgrading of school water, sanitation, and hygiene facilities
- Integration of information and communication technologies

Health

- Construction of health facilities and staff housing
- Integrated community case management through community health workers
- Strengthening of ambulance referral system
- Elimination of user fees
- Introduction of electronic health information systems
- Immunisation campaigns
- Promotion of modern family planning methods and skilled birth attendance
- Comprehensive HIV/AIDS and tuberculosis service delivery
- Bednet distribution
- Community-based rapid diagnostic testing and treatment for malaria

Infrastructure

- Installation and maintenance of water sources
- Promotion and subsidisation of improved sanitation facilities
- Road construction
- Extension of power grids and solar energy
- Expansion of mobile and internet connectivity

up or were discontinued because of funding constraints or regional conflict. In each of the ten scaled up sites, project resources were initially concentrated in a core area of neighbouring villages, referred to as the MV1, and were later extended to additional villages, known as the MV2, as additional resources became available. The MV2s received less intensive interventions than did the MV1s. Together, an MV1 and an MV2 constituted an MV site, each with a population of 25 000–80 000 inhabitants. The MV1s had roughly 6000 inhabitants.

Previous articles about the project's evaluation left unanswered questions and critiques.^{7,10–14} The project's first

public report estimated before–after differences in outcomes in the MV1s, mistakenly referring to these as impacts.⁷ Remans and colleagues¹³ estimated stunting trends in the MV1s and compared them with national trends during a different time period in which stunting remained largely unchanged. For the three countries (Ghana, Nigeria, and Kenya) with both MVP and Demographic and Health Survey (DHS) data, Clemens and Demombynes^{10,15} found somewhat smaller effects on stunting than did Remans and colleagues using a difference-in-differences analysis with a comparison group matched on time period, rural classification, and region.

Pronyk and colleagues¹² compared the MVIs with both national rural areas and sites matched on variables beyond rural classification, although their article did not clearly describe the set of variables or matching procedure used. After comments from Bump and colleagues,¹⁶ Pronyk¹⁷ submitted corrections, withdrawing the comparison with national rural trends in child mortality because of the mismatch of time periods and recalculating the annual decline in child mortality in the MVIs.

Rigorous impact evaluations have not been done for many village-level, integrated, rural development interventions.¹⁰ However, the Southwest Project in China was thoughtfully evaluated.¹⁸ As in our study, evaluation of the Southwest Project faced challenges of reduced statistical power due to village-level interventions and concerns about bias due to the absence of random assignment to the project. By contrast with our study, the investigators in the Southwest Project collected extensive baseline data in non-project areas that could be used to construct a comparison group. Similarly, in an independent evaluation of an MV site in northern Ghana,^{19,20} baseline data were collected in non-project areas and preliminary 2010 census data were accessed to select a comparison group. Abadie and colleagues²¹ described the more general challenge of evaluating interventions assigned (non-randomly) at the aggregate level. They constructed a synthetic comparison using extensive pre-intervention data, which were not readily available for our study. Instead, we analysed available pre-intervention data from a variety of sources and used them in the matching procedure to select villages to form a comparison group.

This endline evaluation addresses three questions: first, in 2015, what were the project's effects on the outcomes of interest within the MVIs? Second, in 2015, were the MDGs and other project-specific targets met within the MVIs? Third, how much was spent (by site, stakeholder, sector, and year) in the MVIs?

Methods

Study design and setting

This evaluation was restricted to the ten scaled up sites (figure 1), focusing only on their core intervention areas (MVIs). For these areas, we estimated the project's impact, target attainment, and on-site spending. To estimate impacts, we took into account previous critiques of the project's evaluation and designed a retrospective, observational study that matched the project villages to comparison villages. The design was completed before endline data collection, and a peer-reviewed evaluation protocol was registered with *The Lancet*.

This study received ethical approval from the institutional review boards at Columbia University, (New York, NY, USA), Ministère de la Santé et de l'Action Sociale (Dakar, Senegal), Ministère de la Santé (Bamako, Mali), Kwame Nkrumah University of Science and Technology (Kumasi, Ghana), University of Ibadan (Ibadan, Nigeria), Mekelle University (Mekelle, Ethiopia), Mbarara University of



Figure 1: Locations of the ten Millennium Villages covered in this evaluation

Science and Technology (Mbarara, Uganda), Kenya Medical Research Institute (Nairobi, Kenya), Ministry of Finance and Economic Planning (Kigali, Rwanda), National Institute for Medical Research (Dar es Salaam, Tanzania), and Ministry of Health (Lilongwe, Malawi).

Outcomes

We chose 40 outcomes of interest from the categories of poverty, agriculture, nutrition, education, child health, maternal health, HIV and malaria, and water and sanitation (table 1). These outcomes were comprised of a subset of MDG indicators and proxies, as well as project-specified outcomes.²² Outcomes and their targets are shown in table 1 and are described in detail in the appendix. We present summary indices for each outcome category. As described in our protocol, we did not separately categorise what we refer to as outcomes into what the evaluation literature^{23,24} refers to as output, outcome, or impact indicators. To supplement the indices, each outcome is presented separately for more nuanced interpretation.

Impact evaluation design

We defined the impact of the MVP as the difference between measured outcomes in the MVIs and what those outcomes would have been in the MVIs without the project. We estimated this difference by comparing outcomes in the MVIs with outcomes in a set of villages where the project was not implemented. Such a comparison was not part of the project design in 2005, which instead focused on target attainment in the MVIs. This decision was based in part on logistical, financial, and ethical complexities related to doing a controlled experiment at the village level. Thus, we used statistical

For the protocol see <http://www.thelancet.com/doi/story/10.1016/html.2015.07.03.2167>

	Target
Proportion of population below 1.25 US\$ (2005 purchasing power parity) per day*	Reduce to 50% of the proportion in 1990
Poverty gap ratio*	Reduce to 50% of the level in 1990
Asset index	No target
Proportion of households that own at least one mobile phone	≥80%
Proportion of farming households that use mineral fertiliser	≥80%
Proportion of farming households that use improved seeds	≥80%
Proportion of children younger than 5 years who are moderately or severely underweight	Reduce to 50% of the proportion in 1990
Proportion of children younger than 5 years who are moderately or severely stunted	Reduce to 50% of the proportion in 1990
Proportion of children younger than 5 years who are moderately or severely wasted	Reduce to 50% of the proportion in 1990
Proportion of children younger than 5 years who tested positive for anaemia	No target
Proportion of school-aged children who tested positive for anaemia	No target
Proportion of women who tested positive for anaemia	No target
Proportion of men who tested positive for anaemia	No target
Adjusted net attendance ratio in primary education	≥90%
Gross attendance ratio for primary education	≥90%
Proportion of pupils starting first grade who reach last grade of primary education	≥90%
Gender parity in primary education	0.97–1.03
Net attendance ratio for preschool	≥90%
Net intake rate for the first grade of primary school	≥90%
Under-5 mortality (per 1000 livebirths)	Reduce to 33% of the level in 1990
Infant mortality (per 1000 livebirths)	Reduce to 33% of the level in 1990
Proportion of children aged 1 year immunised against measles	≥90%
Proportion of children younger than 6 months who are exclusively breastfed	≥50%
Proportion of births attended by skilled personnel	Reduce proportion of unattended births to 25% of the level in 1990
Contraceptive prevalence (any method)	25% nominal increase from level in 1990
Contraceptive prevalence (modern method)	25% nominal increase from level in 1990
Antenatal care coverage, at least one visit with a skilled provider	≥80%
Antenatal care coverage, at least four visits with any provider	≥80%
Proportion of pregnant women tested for HIV during pregnancy	≥90%
Proportion of population aged 15–49 years with comprehensive correct knowledge of HIV/AIDS	≥90%
Proportion of children younger than 5 years who tested positive for malaria	No target
Proportion of school-aged children who tested positive for malaria	No target
Proportion of women who tested positive for malaria	No target
Proportion of men who tested positive for malaria	No target
Proportion of children younger than 5 years who slept under a bednet the previous night	≥80%
Proportion of pregnant women who slept under a bednet the previous night	≥80%
Proportion of households with at least one bednet	≥90%
Proportion of people who used a bednet correctly the previous night	≥90%
Access to improved drinking water source	Reduce proportion without access to 50% of the level in 1990
Access to improved sanitation facility	Reduce proportion without access to 50% of the level in 1990

*Based on consumption.

Table 1: Outcomes and their targets

methods to retrospectively select matched comparison villages and then collected endline data in both the MV1s and the comparison villages. No data were collected by the project in the comparison villages before 2015. Design choices, including matching and sampling methods, were made before endline data collection.²⁵

For each country, we selected comparison villages that best matched its MV1 on several variables measured before project implementation (eg, health measures

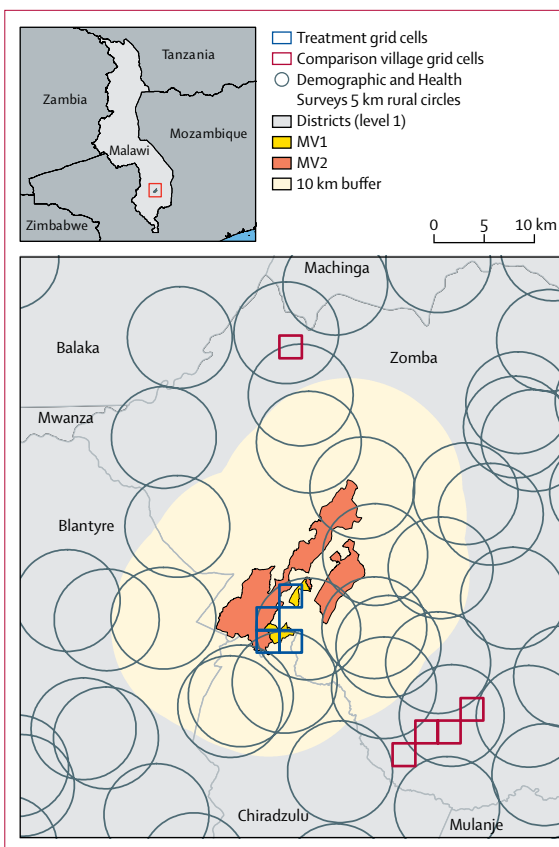
in 2006 and earlier) and variables not affected by the MVP (eg, temperature).²⁶ In our design analysis,²⁷ larger numbers of comparison villages increased statistical power, but with diminishing returns as the number surpassed the number of villages in the MV1. Data collection in each comparison village required permission from village leaders and costly transportation to the area. Balancing these considerations, we chose five comparison villages per country.

To obtain variables on which to match villages, we gathered data from geographic information system (GIS) databases, including agroecological zone, travel time to nearest city with a population of more than 100 000 people, soil composition, vegetation index, temperature, elevation, and population density (appendix).^{28–34} The datasets were stored in the World Geodetic System 1984³⁵ and processed and merged with ArcGIS Desktop version 10.2.2 to obtain values for grid cells equal in area to each country's MV1 (figure 2; appendix). Grid cells overlapping the MV1 (treatment grid cells) were matched to comparison grid cells. We excluded any possible comparison grid cells that overlapped a 10 km buffer around the MV site to minimise the possibility that comparison villages were affected by the project.

We did not use national census data because of its insufficiently precise georeferencing. We also did not use baseline data collected by the MVP because of quality concerns. Instead, we used DHS data, which included many of our outcomes of interest. For each of their sampled areas, DHS provide Global Positioning System coordinates that are displaced by up to 5 km for anonymity.^{36,37} DHS collect data from 20 to 30 households within each sampled area, represented by circles with 5 km radii (figure 2). We used DHS data from 2006 and earlier (appendix). For seven countries (all but Nigeria, Ethiopia, and Tanzania), DHS circles overlapped treatment grid cells. Assuming spatial smoothness, we considered data from any treatment grid cells to be informative of that country's MV1. For these countries, we restricted candidate matches to grid cells overlapping DHS circles.^{38,39} For Nigeria, Ethiopia, and Tanzania, we used only geographical data in the matching.

We matched exactly on country and agroecological zone. For each country, we limited matched comparison grid cells to the MV district and any districts that bordered the MV site. We constrained at least two of the five matches to be within district.⁴⁰ With few treatment grid cells per country, it was difficult to estimate propensity scores, so instead we matched on indices of related variables, which we refer to as matching indices.⁴¹ As a measure of wealth, we used the DHS household asset index, the first principal component of a list of assets.^{42,43} Additionally, we created matching indices for both education and health. Since DHS data include only 20–30 households per area, we fit small-area models using geographical data to improve our estimates of the wealth, education, and health matching indices (appendix).^{44–48}

After restricting by districts and agroecological zone (and, for seven countries, to grid cells with DHS data), we considered sets of five comparison grid cells. To select matches, we estimated a measure of variable imbalance that combined the standardised difference in means and the ratio of standard deviations between the treatment and comparison villages (appendix).^{49,50} To assess the quality of the matches, we presented plots to subject-matter experts (appendix). If concerns were raised about



For more on ArcGIS Desktop see <http://desktop.arcgis.com/en>

Figure 2: Matching results for Malawi

	Number of villages in MV1	Range of MV1 village sizes*	Number of villages in comparison group	Range of comparison village sizes*
Potou, Senegal	14	13–144	9†	10–72
Tiby, Mali	8	40–320	5	45–211
Bonsaaso, Ghana	11	13–258	5	25–446
Pampaida, Nigeria	28	10–94	5	118–507
Koraro, Ethiopia	9	77–239	5	283–545
Ruhiira, Uganda	9	80–192	5	97–153
Sauri, Kenya	11	49–173	5	74–244
Mayange, Rwanda	9	123–283	5	92–261
Mbola, Tanzania	3	235–382	5	325–967
Mwandama, Malawi	5	113–290	5	79–573

*Numbers of households are shown. †The comparison villages in four of Senegal's matched grid cells were too small to allow for a sufficient sample size, so we randomly sampled an additional village in each of those cells.

Table 2: Description of the ten core intervention areas (MV1s) and their comparison villages

the match on particular variables, we assigned those variables increased weight and reran the matching procedure, achieving a closer match on them. Matching results for all sites are shown in the appendix.

Field teams listed all villages within each matched comparison grid cell. We removed any villages outside the range of that country's MV1 village sizes (table 2). Among the remaining villages, we randomly selected one village

For more on **The Registry for International Development Impact Evaluations** see <http://ridie.3ieimpact.org>

per grid cell using a future Nasdaq index as a random seed that was pre-registered with The Registry for International Development Impact Evaluations. If a village's size was unknown, we kept it as a possible selection. If no villages were within the range of the MV1 village sizes, we chose the village whose size was closest to that range. The three countries matched without use of DHS data (Nigeria, Ethiopia, and Tanzania) had the greatest imbalances in the sizes of MV1s versus comparison villages.

Geographical datasets were created in ArcGIS version 10.2.2, with scripts written using Python 2.7.8 in the ArcPy site package Spatial Analyst module. Maps were created with ArcGIS version 10.4. The matching procedure was done with R version 3.1.3.

Target attainment

We defined target attainment as the difference between project outcomes and prespecified targets for these outcomes. Targets were based on official UN MDG targets; international standards; and, where no official UN target or international standard existed, goals set by the MVP sector leaders. 31 of the 40 outcomes were assigned targets. Targets per MV1 are shown in the appendix.

Survey data collection

Throughout the implementation of the MVP, population-based cross-sectional surveys were done every 2 years in the MV1s. At the end of the project in 2015, we collected representative data from cross-sectional surveys in four modules: household surveys, adult surveys, blood testing, and anthropometric measurements. Within each MV1 and comparison village, we used a two-stage design: households were sampled in stage 1, followed by people within households in stage 2.^{51,52} The first stage involved equal-probability simple random sampling of 300 households in each MV1 and comparison group.⁵³ In the second stage, all survey modules except for the household survey used equal-probability systematic sampling from a list of people grouped by household and ordered randomly (conceptually similar to stratifying on household).⁵² The MVP survey tools were adapted in part from the DHS,⁵⁴ UNICEF's Multiple Indicator Cluster Surveys,⁵⁵ and the World Bank's Living Standards Measurement Study surveys⁵⁶ (appendix). Households and individuals were sampled with Stata version 13.0. Survey data were cleaned with the Census and Survey Processing System version 4.1.

Endline data collection was done between March and December, 2015 (see the appendix for the exact dates for each MV1 and its comparison villages). Within each country, data collection was done in the MV1 and its comparison villages within about 2 weeks of each other. Surveys were administered verbally in the local language by module-specialised enumerators after obtaining informed consent. Responses were collected on paper, double-entered electronically, and cleaned. The African Population and Health Research Center did quality

assurance checks of survey data collection in the field and data cleaning.

To handle missing data, we used the *mi* package in R, version 1.0, which iteratively drew imputed values from the conditional distribution for each variable given the observed and imputed values for other variables (appendix). Details about the imputation are described in the appendix, along with proportions of non-responses. Results for an available-case analysis were similar to the imputation-based results presented here.

On-site spending data collection

We estimated on-site spending in the MV1s as expenditures reported by communities, donors, governments, and the MVP (appendix). On-site project spending included the costs of service delivery and on-site management. Off-site project spending, including spending on salaries and overheads for all scientific and support staff at the Earth Institute and Millennium Promise, was excluded. Community, donor, and government on-site spending data were collected through 2014; internal project on-site spending data were collected through 2015. On-site spending is presented as 2005 US\$, adjusted with the US Consumer Price Index.⁵⁷ On-site spending data were analysed with Excel for Mac version 14.0.

Statistical analysis

Project effects and target attainment were estimated with statistical uncertainty and were thus prone to error. When estimating many quantities, the chance of at least one error (eg, an incorrect sign or exaggerated magnitude) increases, creating a problem known as multiple comparisons.⁵⁸⁻⁶⁰ We alleviated this problem by considering countries and outcomes jointly, reducing statistical uncertainty through combining data. We combined outcomes into outcome indices for eight categories (poverty, agriculture, nutrition, education, child health, maternal health, HIV and malaria, and water and sanitation), as well as an overall outcome index. Each village's outcome index is an average of its standardised outcomes within a category. To standardise, we subtracted the mean across villages, divided by the standard deviation across villages, and reoriented so that larger values indicated higher economic development. Project impact and target attainment were estimated for each outcome index. Estimates of project impact and target attainment were also averaged across countries.

We did classical estimation (design-based survey estimation⁶¹) to directly summarise the raw data for each village. Project impact was estimated as average differences in outcomes between project and comparison villages. Target attainment was estimated as average differences between project outcomes and the prespecified targets. We reported separate estimates, with 95% uncertainty intervals (UIs), for each country and outcome, as well as averages across countries and outcomes.

For more on the **Census and Survey Processing System** see <https://www.census.gov/data/software/cspro.html>

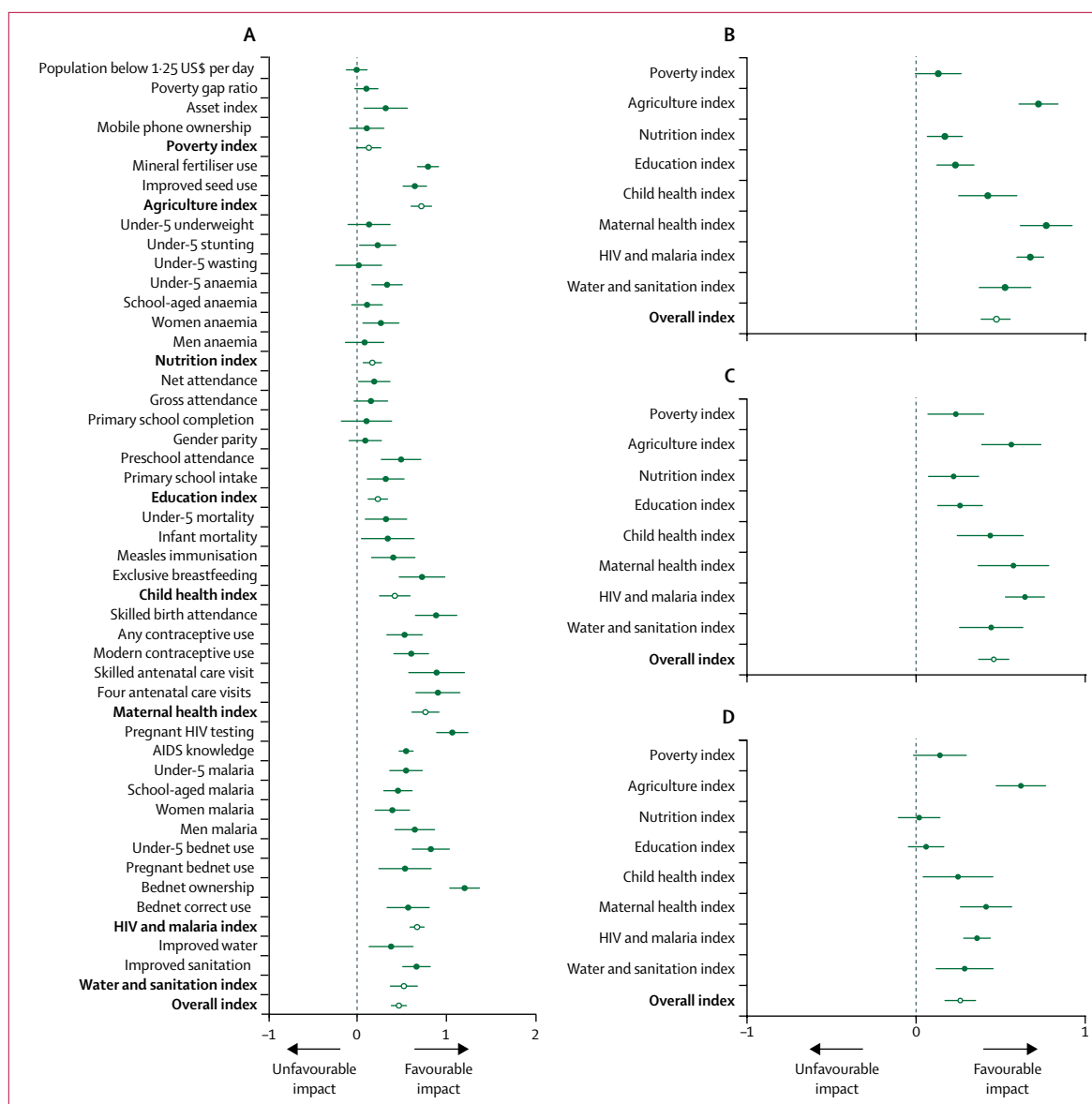


Figure 3: Classical and Bayesian impact estimates for individual outcomes and indices, averaged across countries

Results are on the scale of SDs of the outcomes. Horizontal lines are 95% uncertainty intervals. (A) Classical estimates for individual outcomes and indices. (B) Classical estimates for outcome indices. (C) Bayesian estimates for outcome indices. (D) Classical estimates for outcome indices averaged across only the seven countries matched with Demographic and Health Survey data.

For project impacts, we also did Bayesian analysis to improve on both separate and averaged estimates, fitting a model to obtain country-specific and outcome-index-specific estimates based on information from all countries and outcomes.^{58,60,62} We assessed the model fit via posterior predictive checks (ie, comparing data generated under the model with observed data), which showed that the model fit well (appendix).⁶³ Both classical and Bayesian methods are described in the appendix.

Statistical outcome analyses were done with R version 3.4.3 and Stan version 2.17.2. Statistical graphs were generated in R version 3.4.3.

Role of the funding source

The funders of the evaluation had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

When averaged across the ten MVIs, impact estimates for 30 of the 40 outcomes were significant (95% UIs excluded zero), all of which favoured the MVIs

	MV1	Comparison villages	Impact estimate (95% UI)
Proportion of population below 1.25 US\$ (2005 purchasing power parity) per day	0.60	0.60	0 (-0.04 to 0.04)
Poverty gap ratio	0.26	0.27	-0.02 (-0.04 to 0.01)
Asset index	0.99	0.58	0.41 (0.10 to 0.73)*
Proportion of households that own at least one mobile phone	0.72	0.69	0.02 (-0.02 to 0.06)
Proportion of farming households that use mineral fertiliser	0.66	0.33	0.33 (0.28 to 0.38)*
Proportion of farming households that use improved seeds	0.41	0.19	0.21 (0.17 to 0.25)*
Proportion of children younger than 5 years who are moderately or severely underweight	0.13	0.15	-0.02 (-0.05 to 0.01)
Proportion of children younger than 5 years who are moderately or severely stunted	0.26	0.30	-0.04 (-0.08 to 0.00)*
Proportion of children younger than 5 years who are moderately or severely wasted	0.05	0.05	0.00 (-0.02 to 0.02)
Proportion of children younger than 5 years who tested positive for anaemia	0.32	0.41	-0.09 (-0.14 to -0.04)*
Proportion of school-aged children who tested positive for anaemia	0.41	0.44	-0.03 (-0.08 to 0.02)
Proportion of women who tested positive for anaemia	0.29	0.36	-0.07 (-0.13 to -0.02)*
Proportion of men who tested positive for anaemia	0.24	0.27	-0.02 (-0.08 to 0.04)
Adjusted net attendance ratio in primary education	0.75	0.70	0.05 (0.00 to 0.09)*
Gross attendance ratio for primary education	1.24	1.16	0.08 (-0.02 to 0.17)
Proportion of pupils starting first grade who reach last grade of primary education	0.69	0.66	0.03 (-0.05 to 0.11)
Gender parity in primary education	1.08	1.02	0.06 (-0.06 to 0.19)
Net attendance ratio for preschool	0.44	0.30	0.14 (0.08 to 0.21)*
Net intake rate for the first grade of primary school	0.31	0.22	0.09 (0.03 to 0.14)*
Under-5 mortality (per 1000 livebirths)	62	85	-23 (-40 to -6)*
Infant mortality (per 1000 livebirths)	42	58	-17 (-31 to -2)*
Proportion of children aged 1 years immunised against measles	0.79	0.67	0.12 (0.05 to 0.19)*
Proportion of children younger than 6 months who are exclusively breastfed	0.81	0.57	0.23 (0.15 to 0.32)*
Proportion of births attended by skilled personnel	0.91	0.69	0.22 (0.16 to 0.27)*
Contraceptive prevalence (any method)	0.60	0.46	0.14 (0.09 to 0.19)*
Contraceptive prevalence (modern method)	0.55	0.39	0.16 (0.11 to 0.21)*
Antenatal care coverage, at least one visit with a skilled provider	0.96	0.85	0.12 (0.08 to 0.16)*
Antenatal care coverage, at least four visits with any provider	0.80	0.58	0.22 (0.16 to 0.27)*
Proportion of pregnant women tested for HIV during pregnancy	0.95	0.67	0.28 (0.23 to 0.32)*
Proportion of population aged 15–49 years with comprehensive correct knowledge of HIV/AIDS	0.39	0.23	0.16 (0.14 to 0.19)*
Proportion of children younger than 5 years who tested positive for malaria	0.19	0.34	-0.15 (-0.19 to -0.10)*
Proportion of school-aged children who tested positive for malaria	0.25	0.40	-0.15 (-0.20 to -0.10)*
Proportion of women who tested positive for malaria	0.13	0.22	-0.09 (-0.14 to -0.05)*
Proportion of men who tested positive for malaria	0.11	0.25	-0.14 (-0.19 to -0.09)*
Proportion of children younger than 5 years who slept under a bednet the previous night	0.70	0.49	0.21 (0.16 to 0.26)*
Proportion of pregnant women who slept under a bednet the previous night	0.69	0.49	0.20 (0.09 to 0.31)*
Proportion of households with at least one bednet	0.94	0.68	0.25 (0.22 to 0.29)*
Proportion of people who used a bednet correctly the previous night	0.64	0.53	0.11 (0.06 to 0.15)*
Access to improved drinking water source	0.76	0.64	0.12 (0.04 to 0.19)*
Access to improved sanitation facility	0.63	0.42	0.21 (0.16 to 0.26)*

Data are proportions unless otherwise stated. For a graphical depiction of the results on a standardised scale, see figure 3. UI=uncertainty interval. *Estimated impacts on these outcomes were significant (95% UI excluded zero); they were also favourable.

Table 3: Classical impact estimates averaged across the ten Millennium Village sites, 2015

(figure 3, table 3; appendix). Impacts were particularly substantial in agriculture and health, in which some outcomes were roughly one SD better in the MV1s than in the comparison villages (figure 3). The results were least conclusive for poverty, nutrition, and education outcomes. Compared with estimates made with the classical method (figure 3A, B), the estimated impacts on outcome indices from the Bayesian model, which

avoided over-fitting to each outcome index,^{58,60–63} were slightly closer to the overall estimate (figure 3C).

When averaged across outcomes, Nigeria's MV1 did best relative to its comparison villages, particularly on the maternal health index (figure 4A). The outcome indices of maternal health and HIV and malaria had the largest between-country differences in estimated project impacts. These differences were less pronounced when fitting the

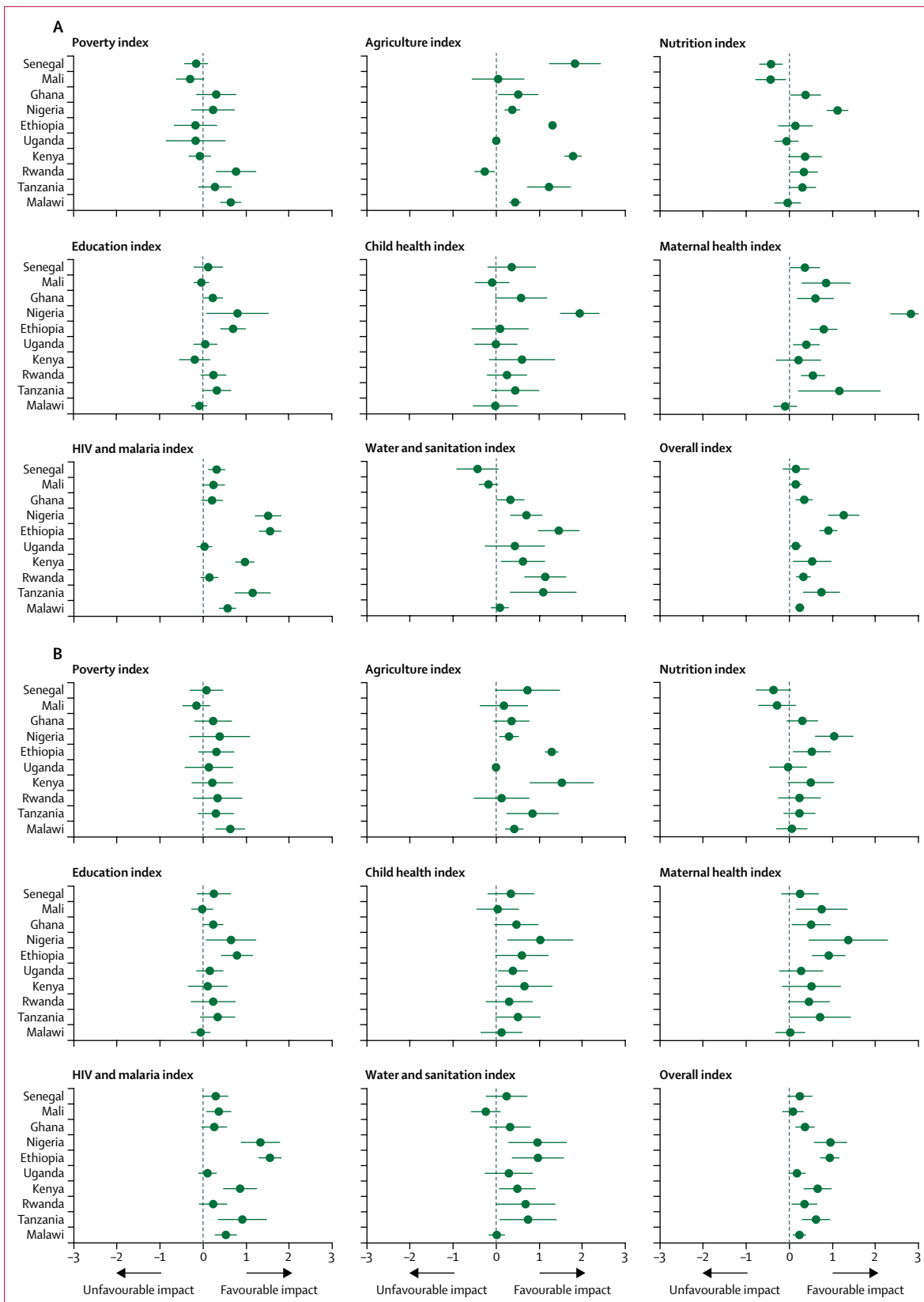


Figure 4: Classical (A) and Bayesian (B) impact estimates for outcome indices, by country
 Results are shown on the scale of SDs of the outcomes. Horizontal lines are uncertainty intervals.

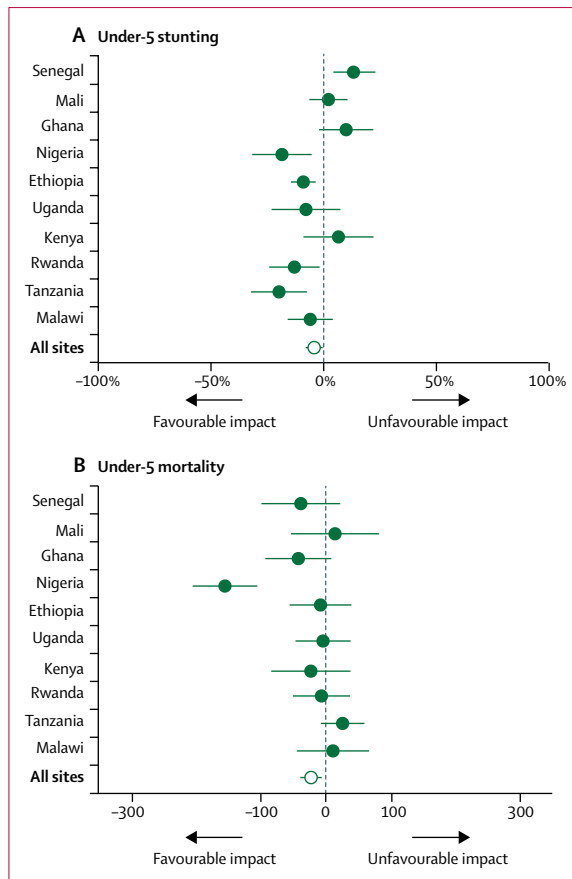


Figure 5: Classical impact estimates for under-5 stunting (A) and under-5 mortality (B)
 Results are proportions for under-5 stunting and deaths per 1000 livebirths for under-5 mortality. Horizontal lines are 95% uncertainty intervals.

Bayesian model (figure 4B), which avoided overfitting to each country.^{58,60–63}

Figure 5 shows country-specific impact estimates for two key outcomes that were the focus of previous MVP evaluations: under-5 stunting and under-5 mortality.^{12,13} Averaged across the ten sites, we estimated that children younger than 5 years were 4% (95% UI 0–8) less likely to be stunted in the MV1s than in the comparison villages. Under-5 mortality was also lower in the MV1s than in the comparison villages by 23 deaths (95% UI 6–40) per 1000 livebirths. For both outcomes, there was substantial site-by-site variation. Country-specific impact estimates for the other 38 outcomes are shown in the appendix.

The sites with the highest estimated overall impacts (Nigeria, Ethiopia, and Tanzania) had comparison villages that were chosen without the use of DHS data because of the absence of data for the matching procedure. Classical estimates of project impact averaged across only the seven countries matched with DHS data are shown in figure 3D.

Figure 6 shows the assessment of target attainment, averaged across countries. Separate results for each

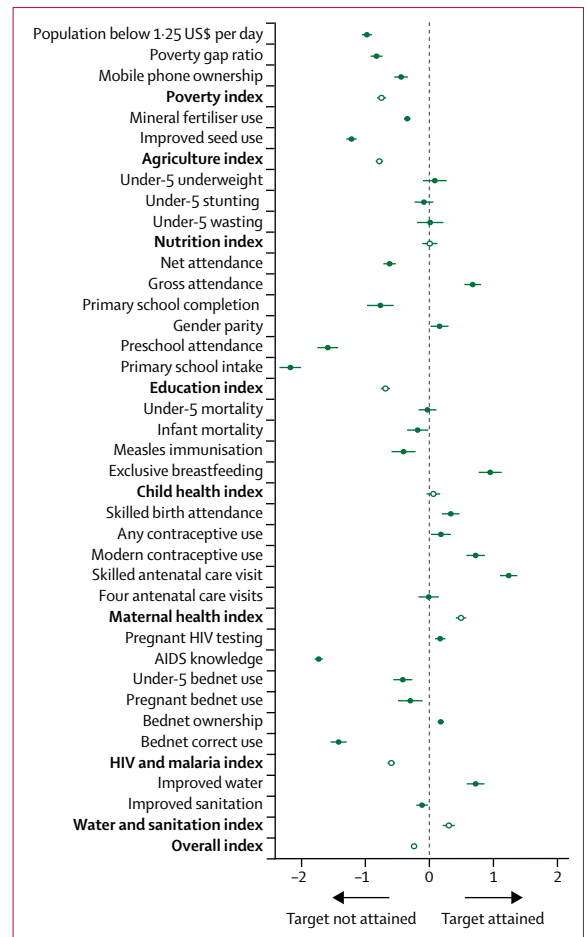


Figure 6: Classical estimates of target attainment, averaged across countries
 Results are shown on the scale of SDs of the outcomes. Horizontal lines are 95% uncertainty intervals.

outcome and country are shown in the appendix. A summary of the impact evaluation and assessment of target attainment, showing the treatment and comparison group estimates relative to the targets, is also shown in the appendix. Averaged across the ten project sites, targets were met for ten of 31 outcomes, targets were not met for 16 outcomes, and target attainment was inconclusive for five outcomes. Maternal health was the best-performing index, with four of the five outcomes meeting their targets.

Figure 7 shows total on-site spending by site; on-site spending by stakeholder, averaged across sites; and project on-site spending by sector, averaged across sites. We observed substantial variation between MV1s in on-site spending (figure 7A). Averaging across the duration of the project (2006–15), five sites (Ghana, Nigeria, Kenya, Rwanda, and Malawi) were above or within the UN Millennium Project recommended range of total on-site spending (\$101–127)⁴ and five (Senegal, Mali, Ethiopia, Uganda, and Tanzania) were below.

Project on-site spending decreased between the first phase (2006–10) and the second phase (2011–15) of the

project because of reduced philanthropic support: on-site spending was, on average, \$66 in the first phase versus \$25 in the second phase (figure 7). This reduction in funding led to an increased emphasis on building systems and capacity of local partners rather than financing of goods and services. Although stakeholders other than the project made up some of the shortfall, the total yearly on-site spending decreased from \$132 in the first phase to \$109 in the second phase (calculated for 2011–14).

In addition to on-site spending data, from the start of the project, the MVP also collected outcome data in the MV1s. The first large, high-quality survey was done in 2010. Differences in estimated outcomes between 2010 and 2015 cross-sections of the MV1s, with 95% UIs, are shown in the appendix. Averaged across the ten project sites, almost all outcomes improved between 2010 and 2015, even as funding was generally reduced. This analysis was not prespecified in the protocol. No project-collected data in the comparison villages were available before 2015, so we could not estimate the project impact on outcomes in 2010 compared with the project impact on outcomes in 2015. Thus, we could not assess whether the impact of the project increased, decreased, or stayed the same over time as on-site spending decreased.

Discussion

This paper summarises analyses of survey and on-site spending data for the endline evaluation of the MVP. The MVP was initiated soon after the UN Millennium Project's recommendations, leading to limitations in the project's design, including the absence of a prospective comparison group for impact evaluation. This study exemplifies methods for retrospective observational studies and addresses the challenge of selecting a comparison group with scarce pre-intervention data.

Averaged across the ten MV1s, the project had a significant and favourable impact on 30 of 40 outcomes of interest and no significant adverse effects. The highest impacts were seen on agriculture, maternal health, and HIV and malaria outcomes. However, substantial variation between sites was observed. The three sites that were unable to be matched with DHS data were estimated to have been most favourably affected by the project, which could indicate poor matching with comparison villages or that these sites truly had the greatest impacts (or somewhere in between). Considering outcomes highlighted in previous evaluations,^{12,13} we found that the MV1s had lower under-5 stunting and mortality estimates than did the comparison villages.

Our indices combined various outcomes along hypothesised causal pathways. For example, bednet ownership, bednet use, and malaria prevalence were grouped together in the HIV and malaria index. For these outcomes, the largest estimated impact was on bednet ownership, which was most directly linked to project activities, whereas the estimated impact on malaria prevalence was less than half in magnitude.

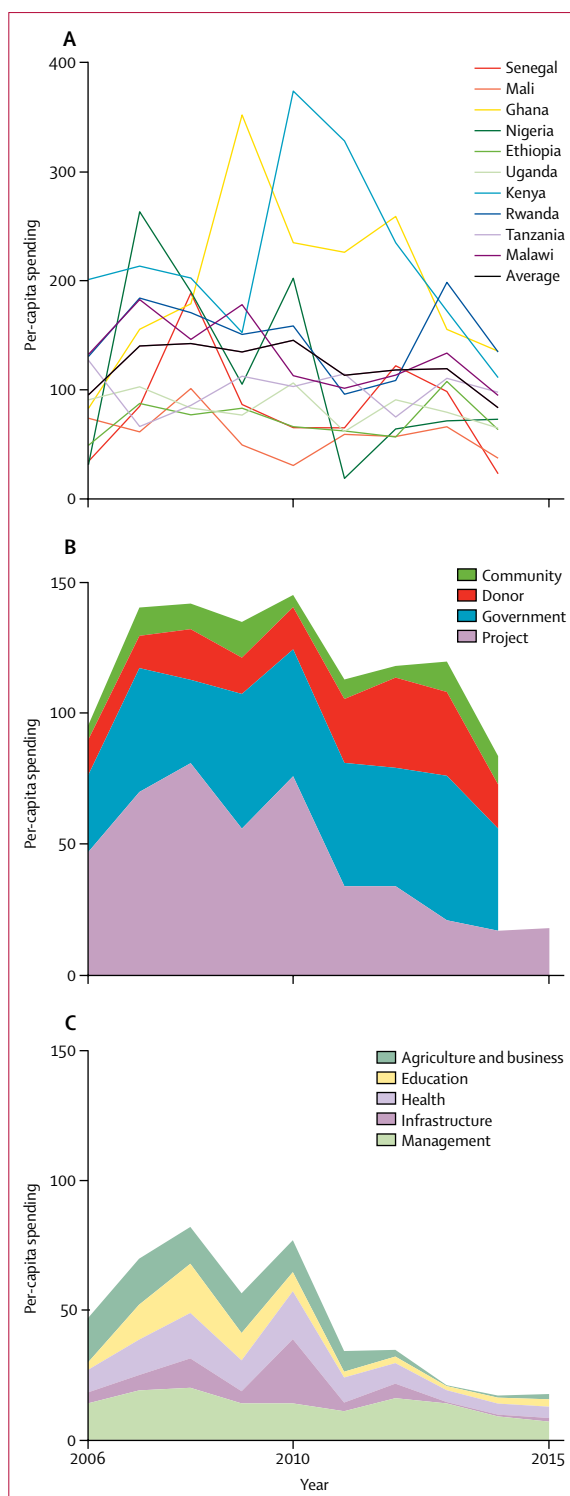


Figure 7: Estimated on-site spending in the MV1s

On-site spending is shown in 2005 US\$, adjusted with the US Consumer Price Index.⁶⁴ (A) Total on-site spending by site. (B) On-site spending by stakeholder, averaged across sites. (C) Project on-site spending by sector, averaged across sites.

According to our assessment of target attainment, averaged across the ten MVIs, roughly a third of the targets were reached. All maternal health targets except one were reached. Some targets were reached across education, child health, HIV and malaria, and water and sanitation indices.

Total on-site spending in the MVIs decreased from an average of \$132 to \$109 between the project's first phase and second phase. Although it would be interesting to see how the impacts of the project changed as on-site spending decreased, we did not estimate impacts over time because we did not have project-collected data in the comparison villages from before 2015. The available trend data in the MVIs between 2010 and 2015 showed that outcomes improved, averaged across the ten sites.

This study has some limitations. Our impact estimates are only interpretable as impacts if two assumptions hold. First, we assumed unconfoundedness—ie, within strata defined by observed variables, the outcomes in the MVIs and comparison villages would have been the same (on average) without the project.^{25,49,58,65–69} Second, we assumed that outcomes in the comparison villages were not affected by the project.⁴⁹

The plausibility of unconfoundedness was limited by the non-random design and scarce data for comparison villages at baseline. Our approach used available data, matching on many variables measured before project implementation or not affected by the MVP.²⁶ Some matching variables were estimated from the DHS sample. We were unable to adjust for any additional confounding by the true (unobserved) variables, although research suggests this residual confounding is not substantial in most cases.^{70,71} For three countries (Nigeria, Ethiopia, and Tanzania), DHS data were not available, so only geographical variables were used. Two unmeasured variables, local political buy-in and community ownership, were not included in the matching, which might have affected the selection of the MV sites and hence were possible confounding variables. We assessed unconfoundedness, and the results did not undermine the assumption's credibility (appendix).

There are several possible routes by which outcomes in the comparison villages could have been affected by the project. First, residents might have migrated between the MV sites and comparison villages. However, at endline evaluation, household heads in the MVIs had lived there for almost 10 years (on average), whereas household heads in the comparison villages had lived in the MV sites for less than 1 year (on average; appendix). Second, residents of comparison villages might have accessed project services, particularly at health facilities. Third, comparison villages might have heard about and adopted MVP interventions. Fourth, outcomes in the MV sites could have affected outcomes in comparison villages—eg, through reduced malaria contagion or sharing of HIV knowledge.^{72,73} Fifth, government spending in the comparison villages

might have been affected by the project as a result of MV sites being targeted or deprioritised for investments. Our matching procedure ensured that comparison villages were at least 10 km away from MV sites in an effort to reduce all but the fifth interference issue.

Generalisability and sustainability are difficult to assess, so extrapolating the results to different scales, locations, and time periods should be done with caution. In particular, even if local political buy-in and community ownership were not confounding variables, they could affect the generalisability of the results.⁷⁴

Our analyses did not take into account spatial correlations (the tendency of closer areas to have more similar outcomes) beyond accounting for clustering into villages and countries.

We did not collect spending data in the comparison villages because of scarce evaluation resources and concerns about the accuracy of data recalled from up to 10 years previously. Knowledge of the difference in spending with and without the project could have enabled a cost-effectiveness analysis.

As in much social science research, both the intervention recipients and the evaluation team were not masked to project assignment. Likewise, recall bias and respondent fatigue might have affected data quality. Although some data were missing because of non-response (appendix), most variables were almost complete, and response rates were similar for MVIs and comparison villages (appendix). Results from an available-case analysis were very similar to the multiple-imputation analysis.

Our non-factorial design did not allow estimation of the effects of component interventions nor their interactions, preventing assessment of the extent of synergistic effects.^{75,76} Similarly, we could not separately estimate the effects of project management from the intervention activities. This paper does not include a process evaluation studying the project's causal pathways.

The MVP did not meet its goal of achieving all of the MDGs, mirroring low attainment of the MDGs across sub-Saharan Africa as a whole.⁶⁴ Both Africa-wide MDG efforts and half of the MVIs received less donor funding than was recommended in the UN Millennium Project's report.^{4,77,78}

The achievements of the MVP in health suggest support for the project's emphasis on strengthening the continuum of care from households, to primary care facilities, and to tertiary care facilities. In particular, we believe that the project's cadres of paid, professionalised community health workers, empowered with smartphones to aid in service delivery and real-time disease monitoring, contributed to the positive results. The project was also an early adopter of interventions and technologies that have since been implemented by development organisations and governments, in part because of the MVP's demonstration and advocacy. These include free mass distribution of insecticide-treated bednets, home-based malaria testing by community health workers using rapid

diagnostic tests, use of mobile health applications for collection of real-time operational data, and micro-grid solar-powered electrification in rural areas. Although poverty is difficult to define and accurately measure, the project's overall positive impact on household asset ownership is a promising indication that living standards were improved.

This impact evaluation was restricted to a cross-sectional, endline comparison with matched villages, using methods specified in the protocol. In the future, additional comparisons would be interesting and useful as sensitivity analyses. For example, comparisons could be made between the MVIs (from data collected by the MVP) and national rural areas (from DHS data). Rural development initiatives, such as those done by the MVP, should be viewed as only one component of an integrated national strategy to end extreme poverty. The MVP was not able to address national-scale infrastructure or systems (such as highways, railways, or supply chains) that are crucial for development in rural and urban areas. However, this endline evaluation might allow some policy implications to be drawn from the project.

Contributors

SMi, AG, and RR contributed to the study design, data analysis, manuscript writing, and figure construction. JC contributed to the study design, manuscript writing, and figure construction. SB contributed to the study design, data collection, and manuscript writing. UKH, MWH, and SES contributed to the study design, data interpretation, and manuscript writing. EAS, AF, and SMa contributed to the study design, data analysis, and manuscript writing. AMZ contributed to the study design and data analysis. LM contributed to the study design and data collection. SO-D and PN contributed to the data collection. CAP contributed to the study design and data interpretation. JDS contributed to the study design, data analysis, data interpretation, manuscript writing, and figure construction. All authors approved the final version to be published.

Declaration of interests

We declare no competing interests.

Acknowledgments

This evaluation was funded by the Open Society Foundations, the Islamic Development Bank, and the governments of Japan, South Korea, Mali, Senegal, and Uganda. The Earth Institute gratefully acknowledges the support of the Bill & Melinda Gates Foundation for supporting MDG-related activities at the Earth Institute and the many funders who contributed to the MVP. We thank the hundreds of people who have contributed to the MVP over the years, particularly the teams in the MV sites, Dakar, Nairobi, and New York. We are also grateful for the contributions of several other groups. The African Population and Health Research Center conducted quality assurance checks of survey data collection in the field and data cleaning. The Stan Development Team provided extensive support with multiple imputation and Bayesian outcome modelling. The Center for International Earth Science Information Network did geographic-data management and mapping, which was key to our study design. We received input from many statisticians and social scientists on study design and analysis. In particular, we thank Macartan Humphreys (WZB Berlin Social Science Center and Columbia University). The Independent Expert Group offered interdisciplinary advice on the evaluation strategy, including questions not addressed in this paper.

References

- UN General Assembly. Resolution 55/2. United Nations Millennium Declaration. New York, NY: United Nations, 2000.
- Sachs JD, McArthur JW. The Millennium Project: a plan for meeting the Millennium Development Goals. *Lancet* 2005; **365**: 347–53.
- Sanchez P, Swaminathan MS, Dobie P, Yuxsel N. Halving hunger: it can be done. New York, NY: UN Millennium Project Task Force on Hunger, 2005.
- UN Millennium Project. Investing in development: a practical plan to achieve the Millennium Development Goals. New York, NY: United Nations Millennium Project, 2005.
- WHO. A summary of the findings of the Commission on Macroeconomics and Health. Geneva: World Health Organization, 2003.
- UN Millennium Project. Q/A on the Millennium Villages, November 2006. https://web.archive.org/web/20071015222205/http://www.unmillenniumproject.org/documents/MV_QA.pdf (accessed Oct 31, 2017).
- The Millennium Villages Project. Harvests of development in rural Africa: the Millennium Villages after three years. New York, NY: The Earth Institute, 2010.
- The Millennium Villages Project. Study protocol. Integrating the delivery of health and development interventions: assessing the impact on child survival in sub-Saharan Africa. New York, NY: The Earth Institute, 2009. <https://ciesin.columbia.edu/confluence/download/attachments/91488269/MVP+Evaluation+Protocol.pdf> (accessed Oct 31, 2017).
- Sanchez P, Palm C, Sachs J, et al. The African Millennium Villages. *Proc Natl Acad Sci USA* 2007; **104**: 1675–80.
- Clemens MA, Demombynes G. When does rigorous impact evaluation make a difference? The case of the Millennium Villages. *J Dev Effect* 2011; **3**: 305–39.
- Clemens MA, Demombynes G. The new transparency in development economics: lessons from the Millennium Villages controversy. Washington, DC: Center for Global Development, 2013. <https://www.cgdev.org/publication/new-transparency-development-economics-lessons-millennium-villages-controversy-working> (accessed Oct 31, 2017).
- Pronyk PM, Muniz M, Nemsler B, et al. The effect of an integrated multisector model for achieving the Millennium Development Goals and improving child survival in rural sub-Saharan Africa: a non-randomised controlled assessment. *Lancet* 2012; **379**: 2179–88.
- Remans R, Pronyk PM, Fanzo JC, et al. Multisector intervention to accelerate reductions in child stunting: an observational study from 9 sub-Saharan African countries. *Am J Clin Nutr* 2011; **94**: 1632–42.
- Wanjala BM, Muradian R. Can big push interventions take small-scale farmers out of poverty? Insights from the Sauri Millennium Village in Kenya. *World Dev* 2013; **45**: 147–60.
- Clemens MA, Demombynes G. Multisector intervention to accelerate reductions in child stunting: an independent critique of scientific method. *Am J Clin Nutr* 2012; **95**: 774–75.
- Bump JB, Clemens MA, Demombynes G, Haddad L. Concerns about the Millennium Villages project report. *Lancet* 2012; **379**: 1945.
- Pronyk P. Errors in a paper on the Millennium Villages project. *Lancet* 2012; **379**: 1946.
- Chen S, Mu R, Ravallion M. Are there lasting impacts of aid to poor areas? *J Public Econ* 2009; **93**: 512–18.
- Masset E, Barnett C, Archarya A, Dogbe T. An impact evaluation design for the Millennium Village project in northern Ghana, 2013. <http://www.ids.ac.uk/publication/impact-evaluation-of-a-new-millennium-village-in-northern-ghana-initial-design-document> (accessed Oct 31, 2017).
- Masset E, Jupp D, Korboe D, et al. Millennium Villages evaluation: midterm summary report. 2016. <http://www.ids.ac.uk/publication/millennium-villages-evaluation-midterm-summary-report> (accessed Oct 31, 2017).
- Abadie A, Diamond A, Hainmueller J. Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. *J Am Stat Assoc* 2010; **105**: 493–505.
- UN Statistics Division. Millennium Development Goal indicators. <http://mdgs.un.org/unsd/mdg/data.aspx> (accessed Feb 11, 2014).
- Vedung E. Public policy and program evaluation. New Brunswick, NJ: Transaction Publishers, 1997.
- Chen H-t. Practical program evaluation. Thousand Oaks, CA: SAGE Publications, 2014.
- Rubin DB. For objective causal inference, design trumps analysis. *Ann Appl Stat* 2008; **2**: 808–40.

- 26 Rosenbaum PR. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J R Stat Soc A* 1984; **147**: 656–66.
- 27 Mitchell S, Ross R, Makela S, et al. Causal inference with small samples and incomplete baseline for the Millennium Villages Project. 2015. http://www.stat.columbia.edu/%7Egelman/research/unpublished/MVP_paper_technical_JRSSA_short.pdf (accessed Aug 13, 2015).
- 28 The CGIAR Consortium for Spatial Information. SRTM 90m digital elevation data. <http://srtm.csi.cgiar.org> (accessed Oct 23, 2014).
- 29 Dixon J, Gulliver A, Gibbon D. Farming systems and poverty: improving farmers' livelihoods in a changing world. Washington, DC: World Bank, 2001.
- 30 European Commission Joint Research Center. Travel time to major cities: a global map of accessibility. <http://forobs.jrc.ec.europa.eu/products/gam/sources.php> (accessed June 13, 2017).
- 31 Global Administrative Areas. GADM database of Global Administrative Areas. Version 2. <http://www.gadm.org> (accessed Feb 1, 2015).
- 32 International Research Institute for Climate and Society. IRI/LDEO climate data library. <http://iridl.ldeo.columbia.edu> (accessed Oct 23, 2014).
- 33 International Soil Reference and Information Centre. Soil property maps of Africa at 1 km resolution. <http://www.isric.org/projects/soil-property-maps-africa-1-km-resolution> (accessed June 13, 2017).
- 34 Socioeconomic Data and Applications Center. Gridded population of the world (GPW), v3. <http://sedac.ciesin.columbia.edu/data/collection/gpw-v3> (accessed Oct 23, 2014).
- 35 National Geospatial-Intelligence Agency. World geodetic system 1984. http://www.unoosa.org/pdf/icg/2012/template/WGS_84.pdf (accessed June 2, 2015).
- 36 The DHS Program. FAQs. <http://dhsprogram.com/faq.cfm> (accessed Feb 23, 2014).
- 37 MEASURE DHS/ICF International. Demographic and Health Survey: sampling and household listing manual. Calverton, MD: ICF International, 2012.
- 38 Cook TD, Shadish WR, Wong VC. Three conditions under which experiments and observational studies produce comparable causal estimates: new findings from within-study comparisons. *J Pol Anal Manag* 2008; **27**: 724–50.
- 39 Steiner PM, Cook TD, Shadish WR, Clark MH. The importance of covariate selection in controlling for selection bias in observational studies. *Psychol Methods* 2010; **15**: 250–67.
- 40 Stuart EA, Rubin DB. Matching with multiple control groups with adjustment for group differences. *J Educ Behav Stat* 2008; **33**: 279–306.
- 41 Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41–55.
- 42 Filmer D, Pritchett LH. Estimating wealth effects without expenditure data—or tears: an application to educational enrollments in states of India. *Demography* 2001; **38**: 115–32.
- 43 Michelson H, Muniz M, DeRosa K. Measuring socio-economic status in the Millennium Villages: the role of asset index choice. *J Dev Stud* 2013; **49**: 917–35.
- 44 Ghosh M, Rao JNK. Small area estimation: an appraisal. *Stat Sci* 1994; **9**: 55–76.
- 45 Ghosh M, Natarajan K. Small area estimation: a Bayesian perspective. In: Ghosh S, Dekker M, eds. *Multivariate analysis, design of experiments and survey sampling*. New York, NY: Wiley, 1999: 69–92.
- 46 Jiang J, Lahiri P. Mixed model prediction and small area estimation. *Test* 2006; **15**: 1–96.
- 47 Nadram B. Bayesian generalized linear models for inference about small areas. In: Rey D, Ghosh SK, Mallick BK, eds. *Generalized linear models*. Boca Raton, FL: CRC Press, 2005: 89–109.
- 48 Rao JNK. *Small area estimation*. Hoboken, NJ: John Wiley and Sons, 2003.
- 49 Imbens GW, Rubin DB. *Causal inference in statistics and social sciences: an introduction*. New York, NY: Cambridge University Press, 2015.
- 50 Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci* 2010; **25**: 1–21.
- 51 Lohr SL. *Sampling: design and analysis*. 2nd edn. Boston, MA: Brooks/Cole, 2010.
- 52 Särndal CE, Swensson B, Wretman J. *Model assisted survey sampling*. New York, NY: Springer-Verlag, 1992.
- 53 Mitchell S, Gelman A, Ross R, et al. Design of the Millennium Villages project sampling plan: a simulation study for a multi-module survey. 2015. <https://arxiv.org/abs/1507.02739> (accessed July 9, 2015).
- 54 The DHS Program. Data tools and manuals. <https://dhsprogram.com/data/Data-Tools-and-Manuals.cfm> (accessed Oct 31, 2017).
- 55 UNICEF Multiple Indicator Cluster Surveys. Surveys. <http://mics.unicef.org/surveys> (accessed Oct 31, 2017).
- 56 World Bank Living Standards Measurement Study. Tools: questionnaires. <http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/EXTLSMS/0,,contentMDK:23636522~pagePK:64168445~piPK:64168309~theSitePK:3358997,00.html> (accessed Oct 31, 2017).
- 57 US Bureau of Labor Statistics. Databases, tables, and calculators by subject. <https://www.bls.gov/data> (accessed Oct 31, 2017).
- 58 Gelman A, Hill JL. *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press, 2007.
- 59 Gelman A, Carlin J. Beyond power calculations to a broader design analysis, prospective or retrospective, using external information. *Psychol Sci* 2014; **9**: 641–51.
- 60 Gelman A, Hill J, Yajima M. Why we (usually) don't have to worry about multiple comparisons. *J Res Educ Eff* 2012; **5**: 189–211.
- 61 Little RJA. To model or not to model? Competing modes of inference for finite population sampling. *J Am Stat Assoc* 2004; **99**: 546–56.
- 62 Gelman A. Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics* 2006; **48**: 432–35.
- 63 Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian data analysis*. 3rd edn. Boca Raton, FL: CRC Press, 2014.
- 64 UN Statistics Division. Millennium Development Goals: 2015 progress chart. 2015. http://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20PC%20final.pdf (accessed May 24, 2017).
- 65 Angrist JD, Pischke JS. *Mostly harmless econometrics: an empiricist's companion*. Princeton, NJ: Princeton University Press, 2009.
- 66 Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005; **61**: 962–72.
- 67 Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Stat Sci* 1999; **14**: 29–46.
- 68 Rubin DB. Inference and missing data. *Biometrika* 1976; **63**: 581–92.
- 69 Rubin DB. Bayesian inference for causal effects: the role of randomization. *Ann Stat* 1978; **6**: 34–58.
- 70 Lenis D, Ebnesajjad CE, Stuart EA. A doubly robust estimator for the average treatment effect in the context of a mean-reverting measurement error. *Biostatistics* 2017; **18**: 325–37.
- 71 Webb-Vargas Y, Rudolph KE, Lenis D, Murakami P, Stuart EA. An imputation-based solution to using mismeasured covariates in propensity score analysis. *Stat Methods Med Res* 2017; **26**: 1824–37.
- 72 Ogburn EL, VanderWeele TJ. Causal diagrams for inference. *Stat Sci* 2014; **29**: 559–78.
- 73 Ogburn EL, VanderWeele TJ. Vaccines, contagion, and social networks. *Ann Appl Stat* 2017; **11**: 919–48.
- 74 Allcott H. Site selection bias in program evaluation. *Q J Econ* 2015; **130**: 1117–65.
- 75 Sachs JD, McArthur JW, Schmidt-Traub C, Kruk M, Bahadur C, McCord G. Ending Africa's poverty trap. *Brookings Pap Econ Act* 2004; **1**: 117–240.
- 76 Sachs JD. Rapid victories against extreme poverty. *Sci Am* 2007; **296**: 34.
- 77 MDG Gap Task Force. *The global partnership for development: the challenge we face*. New York, NY: United Nations, 2013.
- 78 OECD. *Development aid rises again in 2016*. Paris: Organisation for Economic Co-operation and Development, 2017. <https://www.oecd.org/dac/financing-sustainable-development/development-finance-data/ODA-2016-detailed-summary.pdf> (accessed July 14, 2017).