



Using balancing weights to compare performance across facilities providing family planning services in Kenya

Lucas Godoy Garraza¹ · Carolina Cardona² · Peter Gichangi^{3,4,5} · Mary Thiongo^{3,4,5} · Philip Anglewicz² · Leontine Alkema¹

Received: 16 February 2024 / Revised: 7 January 2025 / Accepted: 27 January 2025 /
Published online: 21 February 2025
© The Author(s) 2025

Abstract

Assessing the extent to which the quality of family planning (FP) delivery in facilities makes a difference for key outcomes such as service satisfaction or contraceptive discontinuation is of key interest to the family planning field. However, assessment of this relationship is methodologically challenging due to differences in populations served across facilities. Furthermore, data that connect facilities to the populations served are limited. We use novel data from the Performance Monitoring for Action (PMA) project and a new methodological approach to examine the relationship between facility level characteristics and FP outcomes. The PMA data consist of facility surveys and client exit interviews, and capture women's FP outcomes and include information on characteristics of the individual woman, the facility where the woman obtained her family planning services, and follow-up information on contraceptive use. We use a design-based direct standardization method to balance the distribution of populations served across facilities while controlling for the additional variability induced by the balancing weights. We find significant evidence of variation in FP outcomes across facilities that cannot be accounted for by differences in women characteristics. The type of facility (e.g., dispensary), their size, the proportion of staff present, and whether the facility was public were associated with more positive service satisfaction. A higher ratio of staff to FP visits was predictive of lower contraceptive discontinuation.

Keywords Direct standardization · Weighting · Family planning services performance

✉ Lucas Godoy Garraza
lgodoygarraz@umass.edu

¹ Department of Biostatistics & Epidemiology, University of Massachusetts Amherst, Amherst, USA

² Department of Population, Family and Reproductive Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA

³ Technical University of Mombasa, Mombasa, Kenya

⁴ Department of Public Health and Primary Care, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium

⁵ International Centre for Reproductive Health, Mombasa, Kenya

1 Introduction

Reducing contraceptive discontinuation among women who do not want more children is critical to alleviating the elevated levels of unintended births in Sub-Saharan Africa. The quality of the facilities providing family planning (FP) services could play an important role in these efforts (Cardona et al. 2022; Jain et al. 2019). However, research on the causal link between facility characteristics and FP outcomes has been limited due to data and methodological challenges.

Data on the relationship between family planning outcomes such as contraceptive discontinuation and facility level characteristics have been limited for several reasons. First, studies often use cross-sectional data, like Demographic and Health Surveys (DHS) that measure discontinuation retrospectively (e.g., Ali and Cleland 2010; Bradley et al. 2009). A main limitation in this approach is that factors associated with discontinuation are not measured before the woman stops using, only afterwards. Second, in standard survey designs, there is a lack of information on family planning service delivery characteristics, even though a woman's decision to use contraception may be influenced by characteristics of the health system in her setting such as method availability, distance to the facility, facility type, and quality of care. Moreover, most household or population survey data sources are not structured to enable linkage with patient care data to reveal supply-side dynamics, such as contraceptive stock availability and provider-patient interactions, because they cannot identify the facility where a woman obtained her family planning services.

Data from the Performance Monitoring for Action (PMA) project can be used to address the data limitations. PMA has collected data on both family planning service delivery points, through the PMA facility survey, as well as contraceptive discontinuation among women attending these facilities, through client exit interviews and follow-up surveys. The PMA facility survey includes extensive information on contraceptive stocks, costs, and other related measures in selected facilities. PMA uses a prospective approach to measuring discontinuation among clients visiting the facilities. In this form of data collection, women are interviewed when they receive a method at baseline at the facility, and then followed up four to six months later to see if the woman continued using contraception. This approach to data collection permits the measurement of characteristics at the time when the contraceptive method was acquired that may predict later discontinuation. It also allows for matching individual level discontinuation with facility-level characteristics.

Methodologically, our question of interest is: Does the facility where a woman receives family planning (FP) services makes any difference on service satisfaction or subsequent contraceptive discontinuation? Using observational data to tackle questions such as this one depends on our ability to distinguish different sources of variation and make “fair” comparisons. This endeavor is not trivial because women were not randomly assigned to the facility where they received FP services. Different facilities tend to serve different populations. And a difference in the population served can drive differences in outcomes, making naïve comparison misleading.

Developing measures of performance that adjust for difference in populations served has been central to the literature on “profiling” health care providers, hospitals specially, largely relying on model-based indirect standardization (Normand et al. 1997, 2016). Regression models are used to predict how the population served in each facility would have fared if served in an “average” facility instead. For example, Medicare uses a Bayesian hierarchical model to generate indirectly standardized rates for their “Hospital Compare” (<http://www.medicare.gov/hospitalcompare/>). Comparable regression models have

been used in the FP field to examine the relevance of facility-level characteristics (e.g., Anglewicz et al. 2021).

Direct standardization is an alternative for profiling. It focuses on how each facility would have performed if all of them had served the same population. Traditionally, direct standardization was implemented through stratification and reweighting of the stratum-specific outcome (Keiding and Clayton 2014). Unfortunately, only a limited number of variables can be handled with this model-agnostic approach. Recently, developments from the field of causal inference field have greatly extended the applicability of direct standardization. A notable example is template matching (Silber et al. 2014a, b; Silber et al. 2014a, b), which construct subsamples of individual in each facility with characteristics similar to some target sample using multivariate matching. However, because template matching relies on subsamples, it does require considerable samples as a starting point. For smaller data sets, inverse probability weighting (IPW) methods have been proposed (Keele et al. 2021; Tang et al. 2020). In particular, Keele et al. (2021) propose a procedure to find weights that optimize covariate balance across facilities while controlling for the additional induced variance.

In this study, we apply Keele and colleagues' novel approach of using balancing weights to the problem of examining whether there are differences in outcomes across facilities that are likely due to differences in facility performance, rather than differences in the population served. Further, we used the resulting standardized outcomes as input for a meta-regression to explore which facility-level characteristics are predictive of difference in performance. In the next sections, we introduce data, describe the statistical approach, and discuss findings based on the PMA data from Kenya.

2 Data and measures

2.1 Overview of performance monitoring for action (PMA) data sources

Since 2013, PMA (known from 2013 to 2019 as “PMA2020”) has collected representative data on family planning and contraceptive use in eleven geographies in Africa and Asia. Datasets are publicly available at the PMA website (www.pmadata.org); more information on the study design, sampling approach, and response rates is provided in Zimmerman et al. (2017).

This paper focuses on data collected by the PMA project in Kenya related to facilities providing FP services. To qualify as a “facility”, PMA considered any structure that provided family planning methods or services, ranging from a tertiary hospital to a pharmacy or chemist; and the distribution of these facilities varies across settings. To capture information on facility characteristics, PMA carried out a facility survey. The facilities that PMA selected were those that serve the women and households in the PMA female sample. This includes both public and private facilities, with different sampling approaches for each. For public facilities, PMA selected the primary, secondary, and tertiary facility that serves each enumeration area in the PMA population sample (even if they are not located within the enumeration area). For private facilities, PMA conducted a mapping and listing of all private ones within the enumeration area, and randomly sampled up to three of these facilities. The PMA facility survey includes extensive information on contraceptive stocks, costs, and other related measures, as explained further below. PMA's approach to sampling

facilities is described on its website: <https://www.pmadata.org/media/96/download?attachmentment>

In addition to facility surveys, PMA introduced a novel approach for interviewing clients of facilities, using client exit interviews (CEIs). The CEI was based on visiting the facilities included in the facility survey and selected clients for an exit interview. Specifically, PMA selected facilities where monthly FP client caseloads were at least three per day on average, after which interviewers visited each facility for three days and administered the survey to all women who visited the facility for family planning-related reasons. The CEIs at baseline captured women's characteristics, information related to family planning (FP) behaviors, and women's satisfaction with the facility (explained further below). PMA then followed up with these women six months later and administered a short phone survey that included a measure of whether they continued the method received at baseline. More information on the client exit interviews and facility surveys can be found in Karp et al. (2023), and on the PMA website: <https://www.pmadata.org/data/about-data>.

PMA facility and CEI surveys in Kenya were carried out in November and December of 2020; with follow up phone interviews carried out in July and August of 2021. The sample of clients consisted of 3,663 women of reproductive age who participated of both the baseline and the follow-up interview—the attrition in our sample was 11%. These women were recruited across 395 different facilities. The sample size per facility ranged from 1 to 45 and was smaller than 20 for 93% of the facilities.

2.2 Facility-level data and measures

We used the information collected in the facilities survey to construct measures related to family planning services. We constructed indicator variables to indicate whether the facility offered long-acting reversible contraception (LARC) and short-acting reversible contraception (SARC). LARC methods included implants and intrauterine contraceptive devices (IUD). SARC methods included injectables, contraceptive pill (oral contraceptives), emergency contraception, female and male condoms, diaphragm, contraceptive foam, and standard days method. We also constructed indicator variables to capture information on recent stock outs. Finally, we created covariates related to whether facilities charge fees for family planning services, including an indicator variable to flag whether clients were charged to see a provider for family planning services despite not receiving a method of contraception. All facility characteristics are given in Appendix Table 8.

To overcome data limitations associated with small numbers of CEIs per facility, we introduced a clustering approach to group facilities with similar characteristics. We grouped facilities into clusters with a sample of at least 40 clients yielding a sample of 61 clusters of facilities. The clustering algorithm maximizes differences in certain observed facility level characteristics across clusters. This should increase our ability to detect differences in outcomes associated with these characteristics provided such associations exist. On the other hand, if there is no association between the characteristics used for clustering and the outcome, we expect the clustering to not introduce any bias, since it would amount to clustering at random. In any event, the interpretation of disparities in standardized outcomes as differences in outcomes that cannot be explained by observed differences in the population served remains unaltered. The procedure is explained in detail in the Appendix I. In the remainder of the text, facility-level outcomes refer to average outcomes in the clusters.

2.3 Women's CEIs data and measures of interest

The client exit interviews at baseline capture women's characteristics and satisfaction with the family planning services women received. Follow-up interviews capture information related to contraceptive discontinuation. Women's baseline characteristics are related to marital status, education, births, and wealth, as well as additional information related to family planning (FP) behaviors. An overview of characteristics is given in Table 3. Satisfaction and discontinuation are the primary outcomes of interest.

2.3.1 Satisfaction

We constructed subjective measures of quality of services provided, based on women's satisfaction with the family planning services they received during their visit. Specifically, women reported whether providers and staff at the facility were polite, whether they were satisfied with the service, whether they would refer a relative or a friend to the facility, and whether they would return to the facility. These individual reports were transformed into a binary form and translated into an additive score that ranged from 1 to 5 and aggregated at the facility level.

2.3.2 Discontinuation

We constructed a binary indicator to capture contraceptive discontinuation at follow-up. We measured discontinuation of contraceptive use if a woman reported at follow-up that she is no longer using the contraceptive method provided or prescribed at baseline, she has not switched to an alternative contraceptive method, and she does not intend to become pregnant. In our sample, the outcome was defined for 77% of the women recruited at baseline who, at endline, had not stopped using a contraceptive method with the intent to become pregnant.

3 Methods

3.1 Notation

We observe a sample of $1, \dots, n_j$ women visiting one of the $1, \dots, J$ clusters of facilities. For each woman we observe some outcomes after the visit, denoted by Y_i^O , where superscript O indicates the specific outcome considered. In this study, $O \in \{S, D\}$, where Y_i^S refers to the satisfaction score following the visit and Y_i^D refers to the binary indicator of contraceptive discontinuation at follow-up. The same superscripts are used to denote functions, models and model parameters that are specific to each outcome. For each woman, we also observe a vector of background covariates $X_i \in \mathbb{R}^d$, and indicator Z_i , that denotes cluster membership, with $Z_i = j$ if the woman attended the facility in cluster j . For each cluster of facilities, we observe a vector of facility-level characteristics W_j .

3.2 Estimating standardized facility-level discontinuation rates

Simple comparisons of facility-specific outcomes can be misleading because different facilities serve different populations. We would like to know how the facilities would perform if they served the same set of clients, i.e., a counterfactual question. To estimate this quantity, we used weights that balance the covariate distribution across facilities (Keele et al. 2021).

We make this statement more precise with additional notation. Define the expected value of our outcome given observed covariates x and cluster j as $m_j^O(x) = \mathbb{E}[Y^O|X = x, Z = j]$. The expected overall average outcome in cluster j is $\mu_j^O = \frac{1}{n_j} \sum_{i:Z_i=j} m_j^O(X_i)$. This quantity is not directly comparable across clusters because the distribution of woman-level characteristics is not the same. Thus, the difference between the average outcomes between two clusters reflects both differences in quality of service provided at cluster and differences in the distribution of women attributes.

3.2.1 Target quantity

We aim to produce a target quantity that removes the dependence between the woman characteristics X and the cluster Z . We do this by considering a standardized outcome that takes the expectation of $m_j^O(X_i)$ over a common distribution. While other reference populations are possible, we focus on the empirical distribution of the covariates across all women in the sample (regardless of where they were served),

$$\mu_j^{*O} = \frac{1}{n} \sum_{i=1}^n m_j^O(X_i), \quad (1)$$

where the expected outcome in cluster j for a woman with covariate vector x , $m_j^O(x)$, is computed and averaged over all woman rather than only over those served specifically at cluster j .

3.2.2 Assumptions

For the quantity in Eq (1) to be identifiable we need to assume that, at least in principle, any type of women could receive care at any cluster of facilities (where ‘type’ is defined in terms of X). Formally, $0 < P(Z = j|X = x) < 1$. A full causal interpretation also requires that differences in facility patient mix are fully captured by X or, in other words, that unobserved differences in patient mix do not contribute to the estimates.

3.2.3 Estimation of the target quantity

We follow the approach proposed by Keele et al. 2021 and estimate the average population outcome for cluster j , μ_j^{*O} , with a weighted average of observed outcomes for cluster j , using normalized weights $\hat{\gamma}_i$:

$$\hat{\mu}_j^{*O,W} = \sum_{Z_i=j} \hat{\gamma}_i Y_i^O, \quad (2)$$

with $\sum_{Z_i=j} \hat{\gamma}_i = 1$. The weights are selected to minimize imbalances in covariate distribution by solving the following (convex) optimization problem,

$$\min_{\gamma} \sum_{j=1}^J \left\{ \left\| \bar{X}^{tr} - \sum_{i:Z_i=j} \gamma_i X_i^{tr} \right\|^2 + \lambda n_j \sum_{i:Z_i=j} \gamma_i^2 \right\}, \tag{3}$$

subject to

$$\sum_{i:Z_i=j} \gamma_i = 1, \tag{4}$$

where $\bar{X}^{tr} \equiv \frac{1}{n} \sum_{i=1}^n X_i^{tr}$ and X_i^{tr} is a transformation of the original covariates X_i including standardization and feature expansion. The optimization problem trades off two competing terms for each facility j : to improve balance (and thus reduce bias) versus to keep weights homogeneous (to lower variance introduced by the weighting). For the main analysis we set the penalty very low, prioritizing bias reduction ($\lambda = .001$). We present results from a different choice as a sensitivity analysis ($\lambda = .1$). Additional discussion is included in Appendix II. The covariate distribution is captured through a set of transformed covariates X^{tr} , combining facility-level mean outcomes, tertiles (for continuous covariates), and covariates that indicate membership of specific groups defined by combinations of covariates.

3.3 Examining standardized outcome variation

To determine whether there is evidence of variation in the standardized discontinuation outcome that cannot be accounted for by differences in the distribution of observed individual covariates we use a ‘Q-statistic’ (Hedges and Pigott 2001). The Q-statistic is used in meta-analysis to assess heterogeneity across studies. The Q statistics is given by

$$Q^O = \sum_j \frac{(\hat{\mu}_j^{*O} - \bar{\mu}^O)^2}{(\hat{s}e_j^O)^2 + (\tau^O)^2} \tag{5}$$

where $\bar{\mu}^O = \frac{1}{J} \sum_j \hat{\mu}_j^{*O}$, and $\hat{s}e_j^O$ captures estimation error (i.e., the discrepancy between $\hat{\mu}_j^{*O}$ and μ_j^{*O} , see Appendix II Variance estimation for the weighted estimators). Cross-cluster variation in outcomes beyond estimation error is captured by τ^O . We approximate the permutation distribution of the statistic under the null hypothesis $H_0 : \tau^O = \tau_0^O$ by shuffling Z_i a thousand times and computing balancing weights and Q values at each iteration. This procedure is used for hypothesis testing. We also identify a range of values of τ^O with Q values that would not be rejected by the test for a given level (i.e., with p values larger than, say, α) and among them, the value of τ with the largest p value.¹ Alternatively, model-based estimates of τ are provided in the subsequent analysis.

¹ The procedure is implemented in the package *blkvar* (Miratrix & Pashley 2023)

3.4 Performance drivers

As a final step, we assess the association between the standardized facility-level outcomes and facility-level characteristics. To that end, we introduce a statistical model for the distribution of the standardized outcomes. Specifically, we fit a Bayesian multilevel linear regression model that incorporates three sources of cross-facility variation: variation due to differences in facility-level covariates, variation due to measurement error, and finally, variation across facilities that is not accounted for by the covariates or explained by measurement error. This “meta-regression” approach (Hartung et al. 2008, ch. 10) offers the opportunity to explore which facility-level characteristics are associated with differences in standardized performance.

Specific Bayesian multilevel level regression models differ between satisfaction and discontinuation. We pose the following model for the standardized satisfaction in the j th cluster,

$$\begin{aligned} \hat{\mu}_j^{*W,S} &= \theta_j^S + W_j^T \delta^S + e_j^S, \\ e_j^S | \hat{s}e_j^{W,S} &\sim N\left(0, \left(\hat{s}e_j^{W,S}\right)^2\right), \\ \theta_j^S | \tau &\sim N\left(0, \left(\tau^S\right)^2\right), \end{aligned} \tag{6}$$

where δ is a vector of regression coefficients, relating adjusted performance with the facility level characteristics linearly, θ_j represents variation of performance across clusters not explained by those characteristics, and e_j is the sampling error (the error arising from the observing only a sample of women served in facilities in that cluster). As it is common in meta-analysis or small area estimation, we take the first level variation (i.e., $\hat{s}e_j^{BC,S}$) as a known quantity (its estimation is discussed in Appendix II). The normal distribution for the sampling error can be justified in terms of the expected distribution of the estimator of standardized performance (i.e., a weighted average) on large samples. For discontinuation, this approximation may be poor for small proportion. Therefore, for discontinuation we pose instead

$$\begin{aligned} \hat{V}_j^* &\sim \text{Binomial}\left(n_j^{eff}, \mu_j^{*B}\right), \\ \mu_j^{*B} &= \text{logit}^{-1}\left(\theta_j^B + W_j^T \delta^B\right), \\ \theta_j^B | \tau^B &\sim N\left(0, \left(\tau^B\right)^2\right), \end{aligned} \tag{7}$$

where $\hat{V}_j^* \equiv n_j^{eff} \times \hat{\mu}_j^{*D,W}$, is the “effective” number of cases as in Chen et al. (2014). For Bayesian estimation we need to advance priors for (δ^S, τ^S) and (δ^B, τ^B) , we use flat improper prior for the regression coefficients and weakly informative prior for the variance component (Gelman 2006). Draws from the posterior distribution were obtained via MCMC (Additional details provided Appendix IV).

3.5 Comparison with other weighting strategies

Balancing weights (BW) are constructed to balance the covariate distribution across facilities while controlling the additional variance introduced by the weights. In this approach, the propensity score, i.e., the probability of visiting a particular facility as a function of covariates, is estimated only implicitly (Ben-Michael et al. 2021). An alternative approach starts by explicitly estimating the propensity score, which is then used to construct weights. This approach, known as inverse probability of treatment weights (IPTW), has a longer tradition in epidemiology (Robins et al. 2000) and has more recently been introduced for comparative healthcare quality (Tang et al. 2020). We compare the performance of balancing weights with different versions of IPTW.

To summarize the performance of the weights in terms of reducing the discrepancy in multiple variables we use the Mahalanobis distance. Define $\bar{X}^{tr} \equiv \frac{1}{n} \sum_{i=1}^n X_i^{tr}$ as the average of X_i^{tr} overall, $\bar{X}_j^{tr} \equiv \frac{1}{n_j} \sum_{i:Z_i=j} X_i^{tr}$ as the average in each cluster before weighting, and $\bar{X}_j^{tr,W} \equiv \sum_{i:Z_i=j} \hat{\gamma}_i X_i^{tr}$ as the average in each cluster after weighting. The Mahalanobis distance between covariates for the j th facility before weighting is given by $d_j^T = \sqrt{(\bar{X}^{tr} - \bar{X}_j^{tr})^T \Sigma_X^{-1} (\bar{X}^{tr} - \bar{X}_j^{tr})}$, and the distance after weighting is given by $d_j^W = \sqrt{(\bar{X}^{tr} - \bar{X}_j^{tr,W})^T \Sigma_X^{-1} (\bar{X}^{tr} - \bar{X}_j^{tr,W})}$, where Σ_X^{-1} is the inverse of the variance–covariance matrix of $\left[\bar{X}_j^{tr} \right]_{j:1 \leq j \leq J}$, the matrix of women characteristics. We summarize the change in this measure using the proportional change in the distance, $\Delta_{MH} = 1 - \sum_j d_j^W / \sum_j d_j^T$.

The gains in balance come with a cost in precision. We can assess the changes in precision by computing effective sample sizes. The effective sample size for facility j , is defined as $n_j^{eff} \equiv \frac{(\sum_{i:Z_i=j} \hat{\gamma}_i)^2}{\sum_{i:Z_i=j} \hat{\gamma}_i^2}$, and, provided that $\sum_{i:Z_i=j} \hat{\gamma}_i = 1$, simplifies to $n_j^{eff} = \frac{1}{\sum_{i:Z_i=j} \hat{\gamma}_i^2}$ Potthoff et al. (1992). As is common in survey research, we summarize the difference using the design effect, $Deff = \sum_j n_j / \sum_j n_j^{eff}$.

The IPTW alternatives we consider include estimating the propensity score with multinomial logistic regression and with gradient boosting. Multinomial logistic regression is the most traditional alternative. Because weights developed in this way can be extremely variable, it is a common practice to truncate extreme weights, such as those exceeding a certain percentile (we consider 95 and 75%). Gradient boosting (Friedman 2001) is a machine learning algorithm which approximates the conditional probability function by an ensemble of regression trees. In the version used here (McCaffrey et al. 2013), the hyperparameters are chosen so as to maximize covariate balance.

3.6 Comparison with alternative estimation strategy

A different approach to compare performance across facilities while accounting for differences in the populations served is to directly model the observed outcomes as a function of both individual- and facility-level characteristics using a hierarchical regression model (see

Appendix V). While the approach is typically used as the basis for indirect standardization, it can also be used for (outcome-model-based) direct standardization (Varewyck et al. 2014).

To compare the approaches, we implement a small simulation study introducing a confounding variable. Further, in the application we examine performance drivers using both approaches and compare the facility-level predictors we identify in each case.

4 Simulation study

4.1 Set up

To compare direct standardization with balancing weights with more conventional hierarchical regression we implement a small simulation. The simulation encompasses three scenarios, with different levels of complexity. In all of them, a woman's baseline characteristic is represented disproportionately across facilities. In the more the more benign scenario, however, this baseline covariate is not predictive of the outcome, and, thus, simple average of the outcome in each facility could be used to compare performance without the need of any adjustment. In the second scenario, on the other hand, the baseline covariate is predictive of the outcome, i.e., it is a "confounder". The relationship between the confounder and the outcome, however, is the same across facilities. In the more complex scenario, in contrast, the relationship is inverted in half of the facilities (i.e., the facility is an effect modifier). Other parameters of the simulation are set to resemble the estimation of the discontinuity outcome in our application. The number of facilities and the number of women per facility are taken from the application. The expected discontinuation in each facility is set to approximate the estimated standardized discontinuation in our application. Additional details on the simulation set up are provided in Appendix V.

4.2 Results

Table 1 presents simulation results to compare unweighted, weighted, and hierarchical-model-adjusted approaches. We compute measures of performance for each approach averaged across facilities: the mean absolute error (MAE), the root mean square error (RMSE), and the 95% credible or confidence interval coverage. Expressions for these measures are included in appendix V. The results show that whenever a covariate is related with both the facility and the outcome, the unweighted average is biased. The weights substantially reduce the bias with some loss in precision. When the relationship between the predictor and the outcome is the same across facilities, the hierarchical model reduces the bias about the same extent as the weights without affecting precision. When the relationship between the predictor and the outcome varies across sites (i.e., the outcome model is misspecified), the hierarchical model does not reduce bias to the same extent as weighting does. Coverage of confidence intervals (or credible intervals, in the case of the hierarchical model) does not quite reach nominal levels but is the closest in the case of weight adjustment.

Table 1 Simulated discontinuity for a sample of women visiting different facilities (sample sizes are taken from the real data application)

Scenarios	Measure	Approach		
		Un-weighted	Weight adjusted outcome	Hierarchical model adjusted outcome
A baseline characteristic varies by facility but does not predict the outcome	MAE	0.0026	0.0026	0.0118
	RMSE	0.0307	0.0352	0.0303
	95% CI coverage	0.9344	0.9242	0.8921
A baseline characteristic varies by facility and predicts the outcome	MAE	0.0311	0.0046	0.0160
	RMSE	0.0448	0.0470	0.0324
	95% CI coverage	0.8355	0.9254	0.8990
A baseline characteristic varies by facility and predicts the outcome differently depending on the facility	MAE	0.0312	0.0038	0.0314
	RMSE	0.0497	0.0371	0.0465
	95% CI coverage	0.8632	0.9361	0.8528

The discontinuity in each facility is estimated with unweighted mean, using balancing weights, and using a Bayesian hierarchical generalized linear model. Averages across facilities of mean absolute error (MAE), root mean square error (RMSE) and coverage of 95% CIs, estimated on based on 200 replications. Additional details on the simulation set up are provided in Appendix V

Table 2 Imbalance and effective sample size after weighting. Imbalance (Δ_{MH}) is measured with average Mahalanobis distance

	Balancing weights	IPTW			
		Multinomial logistic regression			Gradient boosting
		Untruncated	Truncated 95%	Truncated 70%	
Relative change in Imbalance (Δ_{MH})	-0.658	1.341	-0.179	-0.469	-0.008
Design effect	2.76	2.77	2.28	1.288	1.522

Design effect measures the loss in precision introduced by the weights

5 Application

5.1 Comparison of different weighting strategies

Table 2 presents the change in imbalance and sample size after weighting. Using untruncated IPW based on multinomial logistic regression to estimate the propensity increases the design effect without reducing imbalances. Implementing the common practice of truncating the weights at the 95% percentile reduces imbalance only to a limited extent. A far larger reduction with comparatively modest increase in design effect would require truncating the weight at 70% percentile. Such level of truncation would be unusual in conventional applications but would offer a substantial reduction in imbalance. Using gradient boosting to estimate the weights did not decrease imbalance in this application. Compared to the alternative weighting strategies, using balancing weights results in the largest reduction of imbalance.

5.2 Sample characteristics before and after weighting

Selected women characteristics are included in Table 3. The table includes a summary measure of the variation of each characteristic across clusters of facilities, the interquartile range (IQR), both before and after the weights are applied. Weighting reduces the IQR for all covariates.

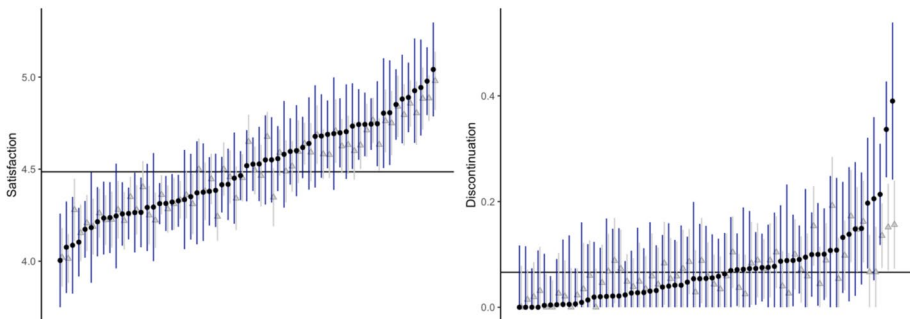
5.3 Variation in outcomes across facilities

Figure 1 shows the standardized facility-level discontinuation rate and level of satisfaction from the family planning services received, obtained after weighting. The difference between the adjusted and unadjusted rates (added in grey in the same figure) are due to weighting. Cross-facility variation in standardized satisfaction and discontinuation is apparent from these figures, but there is also considerable uncertainty around individual estimates and substantial overlap across confidence intervals.

Table 4 summarizes the results of testing the hypothesis that the variation is entirely due to the uncertainty in the estimates. There is considerable evidence of “true”

Table 3 Women characteristics. Overall average in the sample and interquartile range (IQR) across facilities before and after weighting

Variable	Mean	IQR	
		Before weighting	After weighting
Age	28.921	1.344	0.002
Married	0.841	0.091	0.012
<i>Education</i>			
Primary (inc. never attended)	0.409	0.192	0.007
Secondary (or vocational)	0.376	0.130	0.004
College (and University)	0.214	0.168	0.009
Birth events	2.522	0.670	0.006
Household wealth self-rank	4.293	0.750	0.003
This is nearest facility	0.914	0.085	0.006
Health insurance	0.332	0.236	0.004
<i>Contraceptive method before this visit</i>			
No method	0.154	0.101	0.003
Same method	0.568	0.149	0.005
Another method	0.279	0.087	0.004
<i>Type of FP given at baseline</i>			
SARC (rather than LARC)	0.676	0.221	0.006
Single, age 15–25, one kid or none	0.082	0.055	0.020
Married, age 20–30, 3 kids or less	0.430	0.099	0.006
Married, age 31–49, 3 or more kids	0.251	0.114	0.013

**Fig. 1** Standardized facility-level discontinuation rate and satisfaction score (blue), obtained after weighting; raw rates/scores are included in the background (gray)

variation in both satisfaction and discontinuation across facilities that is not accounted by differences in women's characteristics. Specifically, for the standardized estimates, the Q statistics equals 301 and 105 for satisfaction and discontinuation, respectively, unlikely under the null hypothesis of no true variation (p values < 0.01). The standard deviation across clusters, τ , is given by 0.221 (95%CI 0.164, 0.275) in the case of satisfaction and by 0.046 (95%CI 0.023, 0.062) in the case of discontinuation.

Table 4 Average satisfaction score and contraceptive discontinuation rates and estimated standard deviation of the rates and scores across clusters

Outcome Method	Grand Average	Standard deviation (τ)		Q	p value
		Est	95% CI		
<i>Satisfaction</i>					
Unadjusted	4.486	0.218	0.180–0.263	565	0.000
Weighted	4.500	0.212	0.164–0.275	301	0.000
<i>Discontinuation</i>					
Unadjusted	0.066	0.026	0.008–0.047	90	0.009
Weighted	0.067	0.046	0.023–0.062	105	0.007

Confidence intervals (CI) for the standard deviation are obtained by test inversion, the point estimate is the value with the largest p value

Table 5 Comparison of the of average predictive difference (APD) between using direct standardization and a hierarchical model for selected predictors

Outcome Predictor (percentiles 2.5th, 97.5th)	Balancing weights with $\lambda = .001$			Unweighted model-based adjustment		
	APD	SD	P	APD	SD	P
<i>Satisfaction</i>						
Facility type: dispensary (0, 1)	0.209	0.101	0.981*	0.202	0.084	0.989*
Staff here/ total staff (0.29, 0.9)	0.262	0.136	0.972*	0.278	0.118	0.990*
Facility is public (0.66, 1)	−0.537	0.281	0.971*	−0.542	0.254	0.985*
FP visits (90, 644)	−0.270	0.147	0.964*	−0.322	0.125	0.992*
Facility type: hospital (0, 1)	0.189	0.144	0.913	0.269	0.128	0.983*
<i>Discontinuity</i>						
Staff to visits ratio (0.01, 0.63)	−0.123	0.054	0.986*	−0.027	0.031	0.817
Postnatal services (0.6, 1)	−0.009	0.149	0.665	0.060	0.028	0.966*

The APD is the predicted standardized outcome associated with a large change in each predictor holding the remaining predictors constant. A large change in a predictor was defined as the difference between percentiles 2.5th and 97.5th of its distribution. The SD is the posterior standard deviation of APD. The “P” value is the posterior probability of APD having the sign of its point estimate. * is added when the posterior probability is greater than 0.95

It is apparent from Fig. 1 that the amount of variation across facilities, compared to the intra- facility variation, differs between the two standardized outcomes. This feature is frequently summarized by the interunit reliability (IUR), the ratio of the intra-facility variation to the total variation (He et al. 2019a, b; He et al. 2019a, b).² Based

² For a given facility, it is given by $IUR_j = \frac{(\tau^0)^2}{(\hat{\sigma}_j^0)^2 + (\tau^0)^2}$

on the point estimate of τ^O , the median IUR is 0.797 in the case of satisfaction but 0.404 in the case of discontinuation.

5.4 Performance predictors

To summarize the strength of the association between facility-level characteristics and the standardized outcomes, we computed the average predictive difference (APD) associated with a large change in each predictor at a time. Specifically, we calculated the change in average outcomes associated with varying one predictor only, using a range for that predictor based on its 2.5–97.5th percentile (see Appendix V: Average Predictive Difference for further details).

The first three columns of Table 5 show the predictors with APDs with a high probability of being in the direction of the point estimate—above 95% posterior probability. In summary, for satisfaction, higher standardized satisfaction was predicted in dispensaries, and in clusters of facilities with higher proportion of staff present (over the total staff). In contrast, larger facilities, or a larger proportion of public facilities within the cluster, were associated with lower satisfaction. For discontinuation, we find that a lower standardized discontinuity was predicted for clusters of facilities with more staff per visit.

As a byproduct of fitting a Bayesian multilevel model for this analysis, we obtain alternative estimates of typical variation across facilities in the standardized outcome (i.e., τ in Table 2), as well as estimates of this variation after accounting for facility-level predictors. In the case of satisfaction, the across-facility standard deviation decreased from 0.225 (SD=0.0264, similar to the findings in Table 2) before including any predictors to 0.191 (SD=0.0313) after predictors were included. In the case of discontinuation, we estimated an across-facility standard deviation of 0.037 (SD=0.017) before accounting for predictors. This estimate increased to 0.059 (SD=0.026) after predictors were included, indicating the limited predictive power of the covariate set. While unintuitive, this phenomenon can affect multilevel binomial models, due to the fact that the individual-level variability is fixed (Bauer 2009).

The last three columns of Table 5 include results from a conventional unweighted approach (the model for the conventional approach is described Appendix V) to contrast with our analysis, using standardized outcomes. For predicting satisfaction, the unweighted approach results in comparable findings except for the comparison of hospitals to other types of facilities: the unweighted approach suggests that satisfaction is greater in hospitals as compared to other types of facilities. For discontinuation, on the other hand, the two approaches identify different main predictors: the unweighted approach results in a less negative estimate for staff to visit ratios and a positive estimate for provision of postnatal services. Of note, in only 5 of the 61 clusters there are facilities that do not provide postnatal services.

Sensitivity analyses are given in Appendix VI and suggest that findings are robust to different choice of penalty parameter for the weights and updates in modeling assumptions. We used a different distributional assumption for discontinuity's sampling error in the meta-regression and varied the penalty term (to induce less dispersion).

6 Discussion

Despite the importance of assessing the effect of facility-level characteristics on family planning services outcomes like satisfaction and contraceptive discontinuation in places like Kenya, research on this topic is limited due to data availability, measurement, and statistical approaches. We address these limitations here. Firstly, we used newly collected data from the PMA project that allowed for the matching of women with the facilities they attended, and for prospective measurement of discontinuation. Secondly, we implemented a method for direct standardization that allowed us to make comparison of contraceptive discontinuation and satisfaction across facilities as if all the facilities had served women with comparable characteristics. Finally, using Bayesian regression we identified facility-level characteristics that are predictive of differences in standardized discontinuation and satisfaction.

Our analysis provides compelling evidence of heterogeneity in satisfaction and contraceptive discontinuation across facilities that cannot be accounted for by the observed differences in the characteristics of the population served. The level of heterogeneity across facilities was consistent between two different types of analyses (using test statistics and multilevel models). We found various facility-level covariates that were associated with standardized satisfaction. The ratio of staff to visits was the only facility-level covariate identified to have a positive association (with probability greater than 95%) with discontinuation.

Some relationship between facility characteristics, satisfaction and discontinuity identified agree with our hypothesis and prior research (Bellow et al. 2023; Chakraborty et al. 2019; Oyugi et al. 2018). For example, an increased ratio of staff to visits was associated with lower discontinuity. Presumably, the extent to which an adequate workload is achieved affects the quality of services a facility can provide. The size of the facility (as measured by the number of FP visits in the last month) was negatively associated with satisfaction. Larger facilities may be unable to provide more personalized services. Standardized satisfaction was also lower in clusters with larger proportion public facilities. This could be due to public facilities being under resourced. In contrast, the proportion of staff present (out of the total staff) predicted higher satisfaction. This measure might be proxying for staff commitment. The association of satisfaction with dispensaries was more puzzling and might reflect heterogeneity in the population not accounted by the weights, such as particular reasons for the FP visit.

The comparison with alternative weighting procedures shows that in a scenario with a large number of treatments, direct targeting of balance and variance can be beneficial relative to using propensity scores. Simulations illustrate the scenario where weighting adjustment performs better than more conventional hierarchical models, namely when the outcome model is misspecified because the relationship between the confounder and the outcome is not the same across facilities. In our application, our findings were similar between weighting and a more conventional regression model approach in the case of satisfaction, but quite different in the case of discontinuation. This finding implies the need to take caution when aiming to assess variability across facilities, in settings where the population served varies. Our approach improves upon conventional regression model approaches by relaxing the assumption of a linear relation between individual-level characteristics and aggregate outcomes.

Our study is the first to use PMA data to provide estimates of facility-level outcomes using a causal framework. Specifically, we provided the assumptions under which the

differences in standardized outcomes can be causally attributed to differences in the services provided by the facilities. However, a limitation of our study is that the associations between specific facility level characteristics (i.e., the performance “drivers”) and outcomes of interest are not necessarily causal ones. Causal interpretation of the average predictive differences associated with those “drivers” would require additional assumptions.

Taken together, the results of our study highlight the relevance of the facilities in explaining differences in client outcomes—beyond what could be attributed to differences in the population served. Further, they suggest that adequate resources, in particular staffing, may drive those differences.

Appendix

Appendix I. Method for clustering of facilities

Summary

The extent to which direct standardization can improve covariate balance is limited for facilities with small sample size, as is the case in our data. Specifically, in our facility data set, sample sizes per facility range from 1 to 45, with sample size smaller than 20 for 93% of the facilities. We address the issue of small sample sizes by clustering small facilities with similar facility characteristics into larger units of suitable size. Specifically, we grouped facilities into clusters with a sample of at least 40 clients, yielding a sample of 61 clusters of facilities. The characteristics of the resulting clusters, denoted W_j , are a weighted average of the characteristics of the original facilities with weights proportional to the number of follow-up interviews in each facility in the cluster.

Approach

Notation: We denote by n_f the size of the sample of women from the f th facility, for $f = 1, \dots, F$. For each facility, there is auxiliary information on a set of characteristics, $V_f \in \mathbb{R}^p$, such as the type of facility (health center, health clinic, dispensary, pharmacy), whether the facility public, number of staff and indicators on services provided (antenatal, delivery, postnatal, postabortion, HIV related).

Clustering method: There are many ways in which facilities could be grouped to ensure a minimum sample size per cluster. We use a clustering algorithm to identify the grouping that maximizes the internal homogeneity (Assunção et al. 2006), where homogeneity is defined in terms of the set of facility characteristics, V_f , and a robust version of Mahalanobis distance (Rosenbaum 2010, sec 8.3).

Mahalanobis distance is a frequent used dissimilarity measure to summarize differences among multiple features while considering their correlation structure. Given the vector of characteristics, V_f , the Mahalanobis distance between two facilities, f and f' , is defined as

$$d_{f,f'} = \sqrt{(V_f - V_{f'})^T \Sigma^{-1} (V_f - V_{f'})},$$

where Σ is the variance–covariance matrix of V , the F by p matrix of facility characteristics.

While suitable for the multivariate Normal distribution, Mahalanobis distance can have poor performance with long-tailed distributions or rare binary covariates. Rosenbaum (2010) proposed a rank-based modification that avoids those shortcomings. The first part of the modification is to replace the entries of each of the variables, one at a time, by their relative ranking, with average ranks for ties. The second part of the modification is to pre-multiply and post-multiply the covariance matrix of the ranks by a diagonal matrix whose diagonal elements are the ratios of the standard deviation of untied ranks, to the standard deviations of the tied ranks of the covariates. Rosenbaum's modified distance computes the Mahalanobis distance using the relative rankings and the adjusted covariance matrix.

The distance matrix can be further modified to emphasize particular variables using a penalty function. For example, for a particular characteristic w , whenever $|w_f - w_{f'}| > \zeta$, then $d_{f,f'} \leftarrow d_{f,f'} + \beta$, with β some large quantity. This set-up favors near exact matching in the case of categorical variables (Zubizarreta et al. 2011).³ We used this penalty with near exact matching on (i) facility type, (ii) whether the facility is a public facility, and (iii) the location, including the county and whether it a rural or urban setting.

Clustering algorithm: With a suitable measure of distance at hand we could, in principle, consider all assignments of facilities to clusters achieving a predefined minimum sample size per cluster, say n_0 , and chose the one assignment with the minimum within-cluster dissimilarity. Specifically, letting $C_f \in \{1, \dots, K\}$ denote the cluster that the f th facility belongs to, we want to find C_f that minimize within-cluster dissimilarity, $\frac{1}{2} \sum_{k=1}^K \sum_{f:C_f=k} \sum_{f':C_{f'}=k} d_{f,f'}$, subject to $\min \left(\left\{ \sum_{f:C_f=k} n_f \right\}_{k=1}^K \right) \geq n_0$, considering all assignments, C_f , over all possible total number of clusters, $1 < K < F$.⁴ In practice, an exhaustive search is unfeasible, and we need to use some iterative greedy algorithm to approximate the optimum. We used a minimum spanning tree (MST) based algorithm proposed by Assunção et al. (2006). MST algorithms are very fast and has been shown to have competitive performance (Gagolewski et al. 2023).⁵

Once the clusters are found we extend the notation. We index the clusters by $j = 1, \dots, J$, and use n_j to denote the size of the sample of women on the j th cluster. We also have cluster characteristics, denoted by $W_j \in \mathbb{R}^p$, constructed as weighted average of the characteristics of the original facilities, with weights proportional to each facility sample size, e.g., for the g th characteristic, $W_j^g = \frac{1}{n_j} \sum_{f:C_f=j} V_f^g n_f$. We subsequently refer to these clusters of facilities as either clusters or facilities interchangeably.

Appendix II. Weighting to produce standardized outcomes

In this appendix we provide additional details the estimation of the standardized measure of performance.

³ In the case of continues variables, the penalties create so called “calipers” (Zubizarreta et al., 2011).

⁴ Without the minimum sample size constraint, the sum within-cluster dissimilarity would be minimized when $K = F$, i.e., each facility is separately considered a cluster.

⁵ A MST is a connected graph with no circuits, with the minimal sum of the distance over all the edges (Prim 1957). Focusing on the clusters that can be obtained by removing edges of the MST greatly reduces the search space, quickly discarding less promising groupings.

Recall that, for the each of the $i = 1, \dots, n_j$ women served in the j^{th} cluster of facilities have we observe a vector of background characteristics $X_i \in \mathbb{R}^d$ such as age, marital status, or parity. We also observe an outcome, denoted by Y_i^O , where superscript O indicates the specific outcome considered. In this study, $O \in \{S, D\}$, where Y_i^S refers to the satisfaction score following the visit and Y_i^D refers to the binary indicator of contraceptive discontinuation at follow-up. The same superscripts are used to denote functions, models and model parameters that are specific to each outcome. Define $m_j^O(x) \equiv \mathbb{E}(Y^O|X = x, Z = j)$, the conditional expectation of the outcome in cluster j for a woman with a vector of characteristics x .

With this notation, the expected outcome in the j^{th} cluster is given by $\mu_j^O = \frac{1}{n_j} \sum_{i:Z_i=j} m_j^O(X_i)$, which we could easily be estimated with $\bar{Y}_j^O = \frac{1}{n_j} \sum_{i:Z_i=j} Y_i^O$. This quantity, however, is not comparable across clusters. Even if a woman with the same set of characteristics, say x , would have the same expected outcome in cluster j or j' , i.e., even if $m_j^O(x) = m_{j'}^O(x)$, the average outcome in the two clusters could differ because the distribution of covariates is not the same, i.e., $p(x|Z_i = j) \neq p(x|Z_i = j')$. Thus, we are interested in the outcome we would have observed if the population served in all clusters was the same. While other target populations could be defined, we focus on the empirical distribution of the covariates across all facilities, i.e.,

$$\mu_j^{*O} = \frac{1}{n} \sum_i m_j^O(X_i).$$

Weighting

The primary estimator of the quantity of interest is a weighted average of observed outcomes for cluster j , using normalized weights $\hat{\gamma}_i$:

$$\hat{\mu}_j^{*W,O} = \sum_{i:Z_i=j} \hat{\gamma}_i Y_i^O,$$

with $\sum_{Z_i=j} \hat{\gamma}_i = 1$. We want weights that make the covariate distribution of the two populations as similar as possible. Keele et al. (2021) proposed to use the weights that solve the following (convex) optimization problem,

$$\min_{\gamma} \sum_{j=1}^J \left\{ \left\| \bar{X}^{jr} - \sum_{i:Z_i=j} \gamma_i X_i^{jr} \right\|^2 + \lambda n_j \sum_{i:Z_i=j} \gamma_i^2 \right\},$$

subject to

$$\sum_{i:Z_i=j} \gamma_i = 1,$$

$$\ell \leq \gamma_i \leq u,$$

where $\bar{X}^{jr} \equiv \frac{1}{n} \sum_{i=1}^n X_i^{jr}$ and X_i^{jr} is some transformation of the original covariates X_i including standardization and feature expansion. Note that, while feature expansion is “optional”

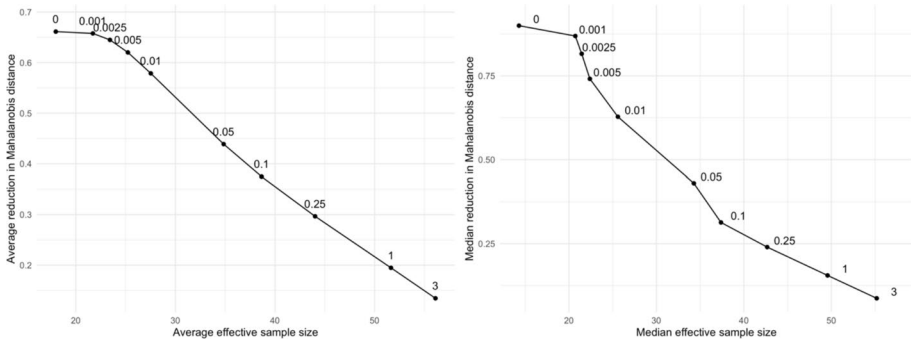


Fig. 2 Covariate Balance vs. Variance Tradeoff as a Function of λ values. Each dot represents the average reduction in Mahalanobis distance (between the mean of the covariates in each cluster and the target population) and the average effective size for differ

the standardization is necessary for the covariates to be given equivalent importance. The optimization problem trades off two competing terms for each facility j : better balance (and thus lower bias) and more homogeneous weights (and thus lower variance); with λ negotiating the tradeoff.

We constraint the weights to be non-negative (equivalently set $\ell = 0$) to avoid extrapolation, but do not set an upper limit. Figure 2 display summaries measure of the tradeoff between covariate balance and variance for several values of λ . Compared with no penalty at all, it appears that setting $\lambda = 0.001$ results on a substantial increase in effective sample size with very little cost in terms of increase imbalance; we selected that value for all the main analyses. As a sensitivity analysis, we also present the results for a choice of $\lambda = 0.1$ that discourage variation in the weights more aggressively (see sensitivity analyses).

Regarding X^{tr} , in addition to the main term for each individual level characteristics, we included indicators for thirdtiles for continuous variables (age, education, self-perceived wealth) to capture nonlinear relationships, and indicators for membership to some large segments based on marital status, age, and parity (i.e., single, between 15 and 25 years old, at most one kid; married, between 20 and 30 years old, at most two kids; married, more than 30 years old, three or more kids) to incorporate important interactions. All features were standardized (subtracted the mean and divided by standard deviation).

The resulting weighted estimator has the benefit of being design-based, i.e., the weights solving the optimization problem do not use any outcome information.

Variance estimation for the weighted estimators

This appendix presents additional detail on how we quantify uncertainty of the weighted and bias-corrected estimators. The approach to quantify uncertainty borrows from the field of survey sampling in which the use of weights has long tradition.

Recall we use Y_i for the individual outcome (either discontinuity or satisfaction), Z_i the binary indicator that individual i is in cluster j , and $\hat{\gamma}_j$ are the weights that minimize the covariate between the sample in cluster j and a target population (the overall sample, in our case). Under the assumption that individual outcomes are sampled independently within facility, the sampling variance, conditional on the weights, is given by

$$\text{Var}\left(\hat{\mu}_j^{*W,O}|\hat{\gamma}_i\right) = \text{Var}\left(\sum_{i:Z_i=j} \hat{\gamma}_i Y_i^O\right) = \sum_{i:Z_i=j} \hat{\gamma}_i^2 \text{Var}(Y_i^O),$$

i.e., a weighted average of the individual-level variance in the outcome, $\text{Var}(Y_i^O)$, which we will denote $(\sigma_j^O)^2$. For each facility, we could estimate the sampling variance (or its square root, the standard error) by plugging in an estimate of the outcome variance, say $\hat{\sigma}_j^O$, i.e.,

$$\widehat{se}\left(\hat{\mu}_j^{*W}|\hat{\gamma}_i\right) = \sqrt{\sum_{i:Z_i=j} \hat{\gamma}_i^2 (\hat{\sigma}_j^O)^2} = \hat{\sigma}_j^O \sqrt{\sum_{i:Z_i=j} \hat{\gamma}_i^2} = \frac{\hat{\sigma}_j^O}{\left(\sum_{i:Z_i=j} \hat{\gamma}_i^2\right)^{-1/2}} = \frac{\hat{\sigma}_j^O}{\sqrt{n_j^{eff}}},$$

where the effective sample size for facility j , is defined as $n_j^{eff} \equiv \frac{(\sum_{i:Z_i=j} \hat{\gamma}_i)^2}{\sum_{i:Z_i=j} \hat{\gamma}_i^2}$, and, provided that $\sum_{z_i=j} \hat{\gamma}_i = 1$, simplifies to $n_j^{eff} = \frac{1}{\sum_{i:Z_i=j} \hat{\gamma}_i^2}$ (Potthoff et al. 1992).

We could estimate the variance of the outcome by its sample counterpart, i.e.,

$$\hat{\sigma}_j^O = \sqrt{\frac{1}{\sum_{i:Z_i=j} \hat{\gamma}_i^2} \sum_{i:Z_i=j} \hat{\gamma}_i^2 \left(Y_{ij}^O - \hat{\mu}_j^{*W,O}\right)^2}.$$

This facility specific estimate, however, may be instable, particularly for facilities with small sample size. For example, in the case of discontinuity, it would be equal to zero in some facilities with no discontinuity events. For this reason, Keele et al. (2021) advocates for the use of a pooled estimate of the variance, a weighted average of the facility-specific estimates, with weights proportional to the effective sample size in each facility, i.e.,

$$\hat{\sigma}_{pool}^O = \sqrt{\frac{1}{\sum_{j=1}^J n_j^{eff}} \sum_{j=1}^J n_j^{eff} (\hat{\sigma}_j^O)^2}.$$

See also Bloom et al. (2017) for a discussion of this approach.

Appendix III. Comparison with other weighting strategies

For comparison purposes, we compute inverse probability of treatment weights (IPTW). ITPW are based on estimates of the (generalized) propensity score, i.e., $e_j(x) \equiv \text{Pr}(Z_i = j|X_i = x)$, where Z_i indicates the facility visited and X_i is the vector of covariates. The IPTW weights are given by

$$\gamma_i^e \equiv \frac{1}{e_{j=Z_i}(x_i)}.$$

We use the normalized version, $\gamma_i^{e'} = \frac{\gamma_i^e}{\sum_{i:Z_i=Z_i} \gamma_i^e}$. Because these weights can be quite variable, it is common practice to trim the weights at a given percentile q , i.e., $\tilde{\gamma}_i^e = \min\left\{\gamma_i^e, \inf\left\{\gamma^e : q \leq \hat{F}(\gamma^e)\right\}\right\}$, before normalization.

As in Tang et al. (2020) we estimate the propensity score using multinomial logistic regression, i.e.,

$$\hat{e}_j(x) = \frac{\exp(X_i^T \beta_j)}{1 + \sum_{k=1}^{J-1} \exp(X_i^T \beta_k)},$$

with $\hat{e}_j(x) = \frac{1}{1 + \sum_{k=1}^{J-1} \exp(X_i^T \beta_k)}$, the baseline category, and X includes a constant. It is common to truncate the weights. We consider weights truncated at several percentiles: 95% to 50% by 5% decrements. We present weights truncated at 95%, as representative of a typical choice, and 70%, because they maximize bias reduction over the grid considered.

As an alternative to multinomial logistic regression, we use a machine learning method to estimate the propensity scores, namely, gradient boosting (Friedman 2001). Gradient boosting algorithms approximate the conditional probability function by an ensemble of regression trees. As in other machine learning approaches, regularization could prevent extreme weights. We use the implementation by McCaffrey et al. (2013), which explicitly incorporates the balancing goal in the estimation, using balance measures as a criteria to select hyperparameters. We note that while this implementation of the algorithm was developed for more than two treatment levels, it is not optimized for the large number of “treatments” in a healthcare comparison setting as in our application.

Appendix IV. Performance “drivers” analysis

To identify facility characteristics that might predict differences in performance we examine the relation between our estimated standardized discontinuation rate, $\hat{\mu}_j^{*W}$, and a set of facility-level characteristics, W_j . This examination is carried out using regression. In the context of meta-analysis, the approach is termed meta-regression (Hartung et al. 2008, ch. 10), and it is used to examine factors driving effect heterogeneity across studies. Similarly, we use in the present context to identify predictors of differences in adjusted performance across clusters. As in the case of meta-regression, even if the input are valid causal estimates (from well conducted RCT), the regression only provides evidence of association.

Specifically, we hypothesize the following model for the standardized satisfaction in the j th cluster,

$$\hat{\mu}_j^{*W,S} = \theta_j^S + W_j^T \delta^S + e_j^S,$$

$$e_j^S | \hat{s}e_j^{W,S} \sim N\left(0, \left[\hat{s}e_j^{S,W}\right]^2\right),$$

$$\theta_j^S | \tau \sim N\left(0, \left[\tau^S\right]^2\right),$$

$$\hat{s}e_j^{W,S} = \frac{\hat{\sigma}_{pool}^W}{\sqrt{n_j^{eff}}},$$

where δ is a vector of regression coefficients, relating adjusted performance with the facility level characteristics linearly, θ_j represents variation of performance across clusters not explained by those characteristics, and e_j is the sampling error (the error arising from the

observing only a sample of women served in facilities in that cluster). As it is common in meta-analysis or small area estimation (in that context, this model is known as Fey-Harriot) we take the first level variation (i.e., \widehat{se}_j) as a known quantity (its estimation is discussed in Appendix II). The normal distribution for the error can be justified in terms of the expected distribution of the estimator of standardized performance (i.e. a weighted average) on large samples.

In the case of discontinuation, we start by defining the “effective number of cases” (as in Chen et al. 2014), $\widehat{V}_j^* \equiv n_j^{eff} \times \widehat{\mu}_j^{*D,W}$. We can then model the effective count as

$$\widehat{V}_j^* \sim \text{Binomial}\left(n_j^{eff}, \mu_j^{*B}\right),$$

$$\mu_j^{*B} = \text{logit}^{-1}\left(\theta_j^B + W_j^T \delta^B\right),$$

$$\theta_j^B | \tau^B \sim N\left(0, [\tau^B]^2\right),$$

Note that we use a different superscript for the standard deviation across facilities, τ^B , which is in different scale than τ^D .

Importantly, with both these models we posit an underlying distribution for the unexplained variation of standardized performance across facilities. The introduction of this random effect induces some extent of pooling across individual estimates of standardized outcomes, particularly when the estimates are more uncertain (e.g., clusters with small effective sample size).

The models are completed with weakly informative prior for the variance across facilities, we use for both outcomes a half student-t with 3 degrees of freedom and a scale parameter of 2.5, i.e., $p(\tau) \propto \left(1 + \frac{1}{b} \left(\frac{\tau}{A}\right)^2\right)^{-\frac{v+1}{2}}$ with $(v, A) = (3, 2.5)$ (Gelman 2006) and flat, improper prior distributions for the vector of regression coefficients, i.e., $p(\delta) \propto 1$. Given the data (i.e., the bias-adjusted estimates taken as data), draws from the posterior distribution were obtained via MCMC. We use the R package brms (Bürkner 2017) to interface with stan (Stan Development Team 2021). We used conventional diagnostics, such as \widehat{R} and effective sample size, to check MCMC convergence.

Appendix V. A comparison with model-based approach to standardization

We compare the results from our analysis from those that would have been obtained from a more conventional, model-based approach to standardization.

An alternative approach to estimate adjusted performance is to model the outcome directly from the start using a hierarchical model. This approach is frequently used as the basis of so called “indirect standardization”. For example, for a binomial outcome, such as contraceptive discontinuation, the likelihood for the outcome would be given by

$$Y_i^D | \eta_i^D \sim \text{Bernoulli}(\text{logit}^{-1}(\eta_i^D)),$$

$$\eta_i^D = \vartheta_{j|i}^D + X_i^T \psi^D,$$

$$\vartheta_j^D | \omega_j^D, \tau_\theta^D \sim N\left(\omega_j^D, [\tau_\theta^D]^2\right),$$

$$\omega_j^D = W_j^T \zeta^D,$$

In turn the satisfaction outcome,

$$Y_i^S | \eta_i^S, \sigma \sim N(\eta_i^S, \sigma^2),$$

$$\eta_i^S = \vartheta_{j|i}^S + X_i^T \psi^S,$$

$$\vartheta_j^S | \omega_j^S, \tau_\theta^S \sim N\left(\omega_j^S, [\tau_\theta^S]^2\right),$$

$$\omega_j^S = W_j^T \zeta^S.$$

where (ψ^D, ζ^D) and (ψ^S, ζ^S) , are vectors of regression coefficients. Bayesian estimation of this model requires prior for all the parameters; flat priors for (ψ^D, ζ^D) and (ψ^S, ζ^S) , for variance parameters, $\tau_\theta^D, \tau_\theta^S$, and σ^2 .

The ϑ 's can be interpreted as an adjusted measure of performance (in the case of a binary outcome, in a logistic scale); for example, if the vector of individual level characteristics, X_i , is centered at their grand mean, then ϑ_j would be the expected value of that measure in the j^{th} facility for a woman with “average characteristics”. Consequently, the ζ 's (as the δ 's in our main approach) capture systematic difference in this measure of adjusted performance associated with facility-level characteristics.

While not a customary for indirect standardization,⁶ we can obtain model-based estimates of our quantities of interest (Varewyck et al. 2014). These are given by,

$$\hat{\mu}_j^{*H,D} = \frac{1}{n} \sum_{i=1}^n \text{logit}^{-1}(\hat{\eta}_i^D(X_i, W_j)),$$

And

$$\hat{\mu}_j^{*H,S} = \frac{1}{n} \sum_{i=1}^n \hat{\eta}_i^S(X_i, W_j),$$

⁶ In the indirect standardization approach the interest lies instead in the outcome that could be expected for the women served in the j^{th} facility if they were served in an “average” facility instead, i.e.,

$$\pi_j^* = \frac{1}{n_j} \sum_{i:Z_i=j}^n \bar{m}(X_i),$$

where $\bar{m}(x) \equiv \mathbb{E}(\mathbb{E}(Y|X = x, Z = j)|X = x)$.

where we use the superscript H, to distinguish these estimates from our weighted (W) alternative. These quantities are computed on each posterior sample from $\hat{\eta}_i^S|Y^S$ and $\hat{\eta}_i^D|Y^D$, respectively, from which we can summarize uncertainty.

Average predictive difference

We can further compare the predictive impact of facility-characteristics implied by this model vis-à-vis the one implied by our proposed approach (see for example, Gelman and Pardoe (2007) for a similar approach). For a given facility-level characteristic, say $u \in W$, we select a pair of values $(u^{(1)}, u^{(2)})$, such as its 2.5% and 97.5% percentiles, and compute the quantity of interest again, altering only that input.

$$PD_{uj}^{O,M} = \frac{1}{n} \sum_{i=1}^n \hat{\mu}_j^{*O,M}(X_i, W_j^{u^{(1)}}) - \frac{1}{n} \sum_{i=1}^n \hat{\mu}_j^{*O,M}(X_i, W_j^{u^{(2)}}),$$

where the superscript $O = \{S, D\}$ stands for outcome (i.e., either satisfaction or discontinuity) and $M = \{W, H\}$ signals the type of estimator (e.g., based on the standardized outcome or on hierarchical regression).

The measure $PD_u^{OM} = \frac{1}{J} \sum_j PD_{uj}^{OM}$, the predictive difference average over the facilities, is directly comparable across approaches.

Simulation set up

To compare direct standardization with balancing weights with more conventional hierarchical regression modeling we implement a simulation.

The simulated data is designed to resemble discontinuation in the real application. We first construct a large population ($N = 10,000$), and then sample repeatedly from this population for each of the $J = 61$ facilities, n_j cases per facility, where n_j is as in the real application.

Each case in the population is characterized by three binary covariates, each drawn from a Bernoulli distribution, i.e., $X_{ki} \sim \text{Bernoulli}(.5)$ for $k = 1, 2, 3$, and J potential outcomes, $Y_i(j)$, drawn from

$$Y_i^*(j)|\eta_{ij} \sim \text{Bernoulli}(\Phi(\eta_{ij})).$$

In each of three scenarios, we set η_{ij} so that the estimand $\mu_j^* = \frac{1}{N} \sum_{i=1}^N \Phi(\eta_{ij})$ is approximately equal to $\hat{\mu}_j^{*W,D}$, the standardized discontinuity estimated in the real application (this is achieved by group centering η_{ij} and adding $\Phi^{-1}(\hat{\mu}_j^{*W,D})$). In all scenarios, we sample repeatedly from this population, with probability proportional to “size”, where the measure of size depends on X_1 and the facility, to ensure unequal distribution. Specifically, we set $s_{ij} = \text{Unif}(0, 2 \cdot \Phi(\eta_{ij})) + \text{Unif}$; in half the facilities the probability of selection is proportional to s_{ij} while in the other half, the probability is proportional to $1/s_{ij}$. In each sample, we set the observe outcome to $Y_i = Y_i^*(z_i)$. The three scenarios differ in the following way:

Scenario 1: No confounders

$$\eta_{ij} = \vartheta_j,$$

where ϑ_j is a facility-specific intercept, set equal to $\Phi^{-1}(\hat{\mu}_j^{*W,D})$ in this scenario.

Scenario 2: A confounder, X_1 , with homogenous effect

$$\eta_{ij} = \vartheta_j + X_{1i},$$

Scenario 3: A confounder, X_1 , which affects the outcome differently, depending on the facility

$$\eta_{ij} = \vartheta_j + X_{1i} \cdot 1(Z_i \leq 30) - X_{1i} \cdot 1(Z_i > 30)$$

To summarize the performance of the different estimators we use the following measures. Let μ^* denote the J -dimension vector of true values, and $\hat{\mu}^*$ the estimator vector. We define the root mean square error as,

$$RMSE(\hat{\mu}^*) \equiv E(\|J^{-1}(\hat{\mu}^* - \mu^*)\|_2) = E\left(\sqrt{\frac{1}{J} \sum_j (\hat{\mu}_j^* - \mu_j^*)^2}\right),$$

and the mean absolute error as,

$$MAE(\hat{\mu}^*) \equiv E(\|J^{-1}(\hat{\mu}^* - \mu^*)\|_1) = E\left(\frac{1}{J} \sum_j |\hat{\mu}_j^* - \mu_j^*|\right).$$

Let $\hat{\mu}_{.95L}^*$ and $\hat{\mu}_{.95U}^*$ be the vectors of lower and upper limits of the 95% confidence or credible intervals around $\hat{\mu}^*$. We define the 95% CI Coverage as,

$$Coverage_{.95}(\hat{\mu}_{.95L}^*, \hat{\mu}_{.95U}^*) \equiv E\left(\frac{1}{J} \sum_j 1(\hat{\mu}_{j,.95L}^* \leq \mu_j^* \leq \hat{\mu}_{j,.95U}^*)\right),$$

where $1(\cdot)$ is the indicator function. These expectations are estimated with the average over 200 replications. For example, $\widehat{RMSE}(\hat{\mu}^*) = \frac{1}{R} \sum_r \|J^{-1}(\hat{\mu}^{*(r)} - \mu^*)\|_2$, where $\hat{\mu}^{*(r)}$ is the value of the estimator vector in the r^{th} replication, and $R = 200$.

Appendix VI. Sensitivity analyses

In this appendix, we present result after altering the balancing weights by increasing the penalty for dispersion.

Altering the choice of penalty to obtain balancing weights.

Table 6 Imbalance and bias reduction and effective sample size after weighting

Measure	Penalty	
	$\lambda = 0.001$	$\lambda = .1$
Percent imbalance reduction	0.658	0.375
Percent bias reduction: discontinuity	0.808	0.527
Percent bias reduction: satisfaction	0.740	0.512
Effective sample size (total sample size 3663)	1325	2358

Imbalance is measured with average Mahalanobis distance “Bias” is estimated by weighting imbalance along different covariates before and after weighting by their ability to predict the outcome based on a linear regression. The effective sample size accounts for the loss in precision induced by the variation in weights

Table 7 Comparison of average predictive difference (APD) using set of weights based on different values for the penalty parameter

Outcome Predictor (percentiles 2.5th, 97.5th)	Balancing weights with $\lambda = .001$			Balancing weights with $\lambda = .1$		
	APD	SD	P	APD	SD	P
<i>Satisfaction</i>						
Facility type: dispensary (0, 1)	0.209	0.101	0.981*	0.195	0.088	0.988*
Staff here/ total staff (0.29, 0.9)	0.262	0.136	0.972*	0.275	0.120	0.985*
Facility is public (0.66, 1)	-0.537	0.281	0.971*	-0.555	0.248	0.983*
FP visits (90, 644)	-0.270	0.147	0.964*	-0.315	0.131	0.990*
<i>Discontinuity</i>						
Staff to visits ratio (0.01, 0.63)	-0.123	0.054	0.986*	-0.087	0.041	0.980*

The APD is the predicted standardized outcome associated with a large change in each predictor holding the rest constant. Large changes in a predictor were defined as changes between percentiles 2.5th and 97.5th of its distribution. The SD is the posterior standard deviation of APD. “P” value is the posterior probability of APD having the sign of its point estimate

Recall that the penalty parameter, λ , regulates the tradeoff between precision (measured for example by the effective sample size) and covariate imbalance (and consequent bias). Since different choices are possible, we rerun the analysis by selecting a different value for lambda, $\lambda = .1$, which penalizes dispersion more aggressively.

Results based on the new set of weights are presented in Table 6. As we penalized the weight variability more, the weights can remove less imbalance. Nevertheless, the results are pretty much unaltered. See Table 7

Appendix VII. Supplementary tables

See Table 8

Table 8 Facility characteristics reference

Variable	Description
providedPERM	Facility provides permanent methods
providedLARC	Facility provides LARC methods
providedSARC	Facility provides SARC methods
stockoutLARC	Out of stock at least once in last 3 months of LARC method
stockoutSARC	Out of stock at least once in last 3 months of SARC methods
chargedPERM	Facility charges for permanent methods
chargedLARC	Facility charges for LARC methods
chargedSARC	Facility charges for SARC methods
Postnatal	Facility provides postnatal services
Antenatal	Facility provides antenatal services
Delivery	Facility provides delivery services
Abortion	Facility provides abortion services
Postabortion	Facility provides postabortion services
HIV services	Facility provides HIV services
Facility type	Health Center; Health Clinic; Dispensary; Pharmacy
Public	Facility is public
fp_visits	Total number of family planning visits (new and continuing) in the last completed month
visits_new_total	The number of new clients who received family planning services in the last completed month/ Total number of family planning visits (new and continuing) in the last completed month
staff_here_total	Staff present the day of the survey/ Staff
staff_visits	Staff present the day of the survey/ The number of new clients who received family planning services in the last completed month
county	County location
rural	Rural or urban location

Appendix VIII. Supplementary figures

See Figs. 3, 4 and 5

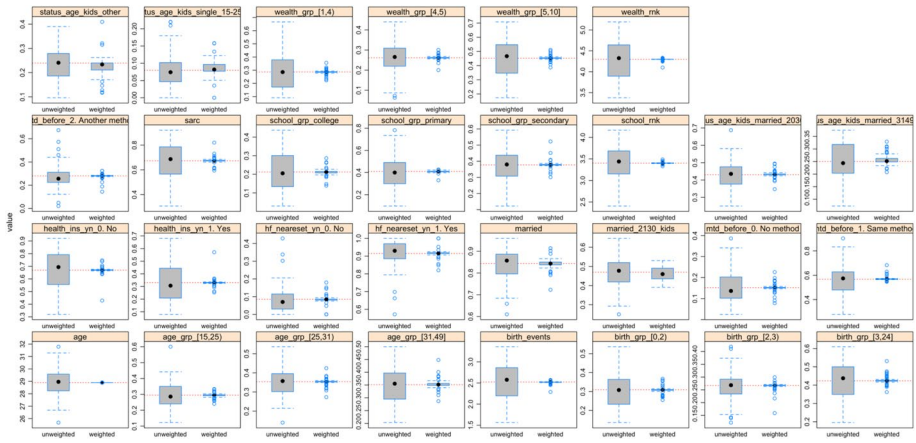


Fig. 3 Boxplots of mean outcomes per cluster before and after weighting. Red line represents the marginal mean, i.e., the average outcome for all women combined

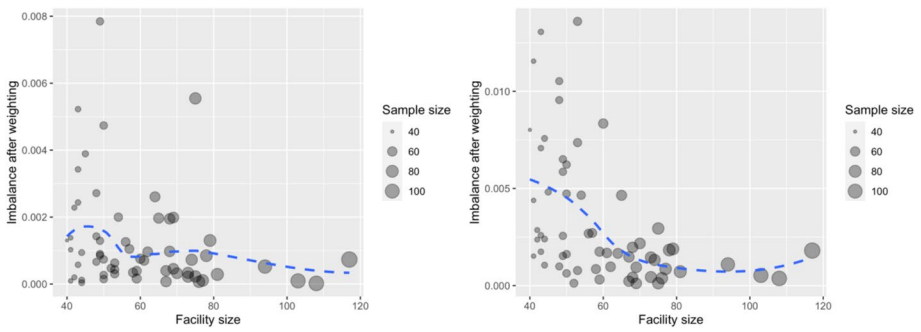


Fig. 4 Residual imbalance after weighting by size. Imbalance is measured with average Mahalanobis distance (d^W) discussed in Appendix II

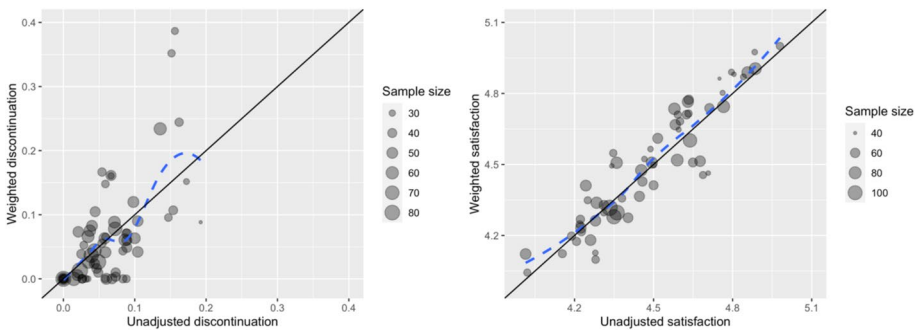


Fig. 5 Unadjusted vs. weighted discontinuation

Acknowledgements We thank Navideh Noori from Institute for Disease Modeling, Bill & Melinda Gates Foundation for her helpful comments

Author contributions L.G.G., C.C., P.A., and L.A conceived the study. L.G.G. designed the analysis with L.A. guidance. L.G.G. implemented the analysis. L.G.G., C.C., P.A. and L.A drafted the manuscript. P.G. and M.T. revised the manuscript critically and contributed to the interpretation of the results.

Funding Funding for this study was provided by the Bill & Melinda Gates Foundation under grant numbers OPP10709004 and INV-00844. Funders were not involved in any aspect of the study design, data collection, and analysis, nor the interpretation and writing of the manuscript.

Data availability Data is publicly available at the PMA website (www.pmadata.org). Analysis code is available at https://github.com/AlkemaLab/pma_facilities.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ali, M.M., Cleland, J.: Contraceptive switching after method-related discontinuation: levels and differentials. *Stud. Fam. Plann.* **41**(2), 129–133 (2010). <https://doi.org/10.1111/j.1728-4465.2010.00234.x>
- Anglewicz, P., Cardona, C., Akinlose, T., Gichangi, P., OlaOlorun, F., Omoluabi, E., Thiogo, M., Akilimali, P., Tsui, A., Kayembe, P., PMA Agile Principal Investigators Group: Service delivery point and individual characteristics associated with the adoption of modern contraceptive: a multi-country longitudinal analysis. *PLoS ONE* **16**(8), e0254775 (2021). <https://doi.org/10.1371/journal.pone.0254775>
- Assunção, R.M., Neves, M.C., Câmara, G., Da Costa Freitas, C.: Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *Int. J. Geogr. Inf. Sci.* **20**(7), 797–811 (2006). <https://doi.org/10.1080/13658810600665111>
- Bauer, D.J.: A note on comparing the estimates of models for cluster-correlated or longitudinal data with binary or ordinal outcomes. *Psychometrika* **74**(1), 97–105 (2009). <https://doi.org/10.1007/s11336-008-9080-1>
- Bellow, N., Dougherty, L., Nai, D., Kassegne, S., Nagbe, R.H.Y., Babogou, L., Guede, K.M., Silva, M.: Improving provider and client communication around family planning in Togo: results from a cross-sectional survey. *PLOS Global Public Health* **3**(6), e0001923 (2023). <https://doi.org/10.1371/journal.pgph.0001923>
- Ben-Michael, E., Feller, A., Hirshberg, D.A., & Zubizarreta, J. R.: The balancing act in causal inference (arXiv:2110.14831) (2021). arXiv. <http://arxiv.org/abs/2110.14831>
- Bloom, H.S., Raudenbush, S.W., Weiss, M.J., Porter, K.: Using multisite experiments to study cross-site variation in treatment effects: a hybrid approach with fixed intercepts and a random treatment coefficient. *J. Res. Educ. Effect.* **10**(4), 817–842 (2017). <https://doi.org/10.1080/19345747.2016.1264518>
- Bradley, S.E., Schwandt, H.M., & Khan, S.M.: Levels, trends, and reasons for contraceptive discontinuation (DHS Analytical Studies No. 20). ICF Macro (2009). <http://dhsprogram.com/pubs/pdf/AS20/AS20.pdf>
- Bürkner, P.-C.: **brms**: An R package for bayesian multilevel models using *Stan*. *J. Stat. Softw.* (2017). <https://doi.org/10.18637/jss.v080.i01>
- Cardona, C., OlaOlorun, F.M., Omulabi, E., Gichangi, P., Thiogo, M., Tsui, A., Anglewicz, P.: The relationship between client dissatisfaction and contraceptive discontinuation among urban

- family planning clients in three sub-Saharan African countries. *PLoS ONE* **17**(8), e0271911 (2022). <https://doi.org/10.1371/journal.pone.0271911>
- Chakraborty, N.M., Chang, K., Bellows, B., Grépin, K.A., Hameed, W., Kalamar, A., Gul, X., Atuyambe, L., Montagu, D.: Association between the quality of contraceptive counseling and method continuation: findings from a prospective cohort study in social franchise clinics in Pakistan and Uganda. *Glob. Health Sci. Pract.* **7**(1), 87–102 (2019). <https://doi.org/10.9745/GHSP-D-18-00407>
- Chen, C., Wakefield, J., Lumely, T.: The use of sampling weights in Bayesian hierarchical models for small area estimation. *Spat. Spatio-Temp. Epidemiol.* **11**, 33–43 (2014). <https://doi.org/10.1016/j.sste.2014.07.002>
- Friedman, J.H.: Greedy function approximation: a gradient boosting machine. *Ann. Stat.* (2001). <https://doi.org/10.1214/aos/1013203451>
- Gagolewski, M., Cena, A., Bartoszuk, M., & Brzozowski, Ł.: Clustering with minimum spanning trees: How good can it be? (arXiv:2303.05679) (2023). arXiv. <http://arxiv.org/abs/2303.05679>
- Gelman, A.: Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1**(3), 515–534 (2006). <https://doi.org/10.1214/06-BA117A>
- Gelman, A., Pardoe, I.: 2. Average Predictive Comparisons for Models with Nonlinearity, Interactions, and Variance Components. *Sociol. Methodol.* **37**(1), 23–51 (2007). <https://doi.org/10.1111/j.1467-9531.2007.00181.x>
- Hartung, J., Knapp, G., Sinha, B.K.: *Statistical Meta-Analysis with Applications*. John Wiley & Sons, Inc., Hoboken (2008)
- He, K., Dahlerus, C., Xia, L., Li, Y., Kalbfleisch, J.D.: The profile inter-unit reliability. *Biometrics* **76**(2), 654–663 (2019a). <https://doi.org/10.1111/biom.13167>
- He, K., Kalbfleisch, J.D., Yang, Y., Fei, Z.: Inter-unit reliability for nonlinear models. *Stat. Med.* **38**(5), 844–854 (2019b). <https://doi.org/10.1002/sim.8005>
- Hedges, L.V., Pigott, T.D.: The power of statistical tests in meta-analysis. *Psychol. Methods* **6**(3), 203–217 (2001). <https://doi.org/10.1037/1082-989X.6.3.203>
- Jain, A., Aruldas, K., Mozumdar, A., Tobey, E., Acharya, R.: Validation of two quality of care measures: results from a longitudinal study of reversible contraceptive users in India. *Stud. Fam. Plann.* **50**(2), 179–193 (2019). <https://doi.org/10.1111/sifp.12093>
- Karp, C., OlaOlorun, F. M., Guiella, G., Gichangi, P., Choi, Y., Anglewicz, P., & Holt, K.: Validation and predictive utility of a person-centered quality of contraceptive counseling (QCC-10) scale in sub-Saharan Africa: a multicountry study of family planning clients and a new indicator for measuring high-quality, rights-based care. *Stud. Fam. Plan.* **54**(1), 119–143 (2023). <https://doi.org/10.1111/sifp.12229>
- Keele, L., Ben-Michael, E., Feller, A., Kelz, R., & Miratrix, L.: Hospital quality risk standardization via approximate balancing weights (arXiv:2007.09056). (2021) arXiv. <http://arxiv.org/abs/2007.09056>
- Keiding, N., Clayton, D.: Standardization and control for confounding in observational studies: a historical perspective. *Stat. Sci.* (2014). <https://doi.org/10.1214/13-STS453>
- McCaffrey, D.F., Griffin, B.A., Almirall, D., Slaughter, M.E., Ramchand, R., Burgette, L.F.: A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat. Med.* **32**(19), 3388–3414 (2013). <https://doi.org/10.1002/sim.5753>
- Miratrix, L., Pashley, N.: blkvar: ATE and treatment variation estimation for blocked and multisite RCTs (Version 0.0.1.5) [Computer software]. (2023)
- Normand, S.-L.T., Glickman, M.E., Gatsonis, C.A.: Statistical methods for profiling providers of medical care: issues and applications. *J. Am. Stat. Assoc.* **92**(439), 803–814 (1997). <https://doi.org/10.1080/01621459.1997.10474036>
- Normand, S.-L.T., Ash, A.S., Fienberg, S.E., Stukel, T.A., Utts, J., Louis, T.A.: League tables for hospital comparisons. *Annu. Rev. Stat. Appl.* **3**(1), 21–50 (2016). <https://doi.org/10.1146/annurev-statistics-022513-115617>
- Oyugi, B., Kioko, U., Kaboro, S.M., Okumu, C., Ogola-Munene, S., Kalsi, S., Thiani, S., Gikonyo, S., Korir, J., Baltazar, B., Ranji, M.: A facility-based study of women’s satisfaction and perceived quality of reproductive and maternal health services in the Kenya output-based approach voucher program. *BMC Pregnancy Childbirth* **18**(1), 310 (2018). <https://doi.org/10.1186/s12884-018-1940-9>
- Potthoff, R.F., Woodbury, M.A., Manton, K.G.: “Equivalent sample size” and “equivalent degrees of freedom” refinements for inference using survey weights under superpopulation models. *J. Am. Stat. Assoc.* **87**(418), 383–396 (1992). <https://doi.org/10.1080/01621459.1992.10475218>
- Prim, R.C.: Shortest connection networks and some generalizations. *Bell Syst. Tech. J.* **36**(6), 1389–1401 (1957). <https://doi.org/10.1002/j.1538-7305.1957.tb01515.x>
- Robins, J.M., Hernán, M.Á., Brumback, B.: Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**(5), 550–560 (2000). <https://doi.org/10.1097/00001648-200009000-00011>
- Rosenbaum, P.R.: *Design of Observational Studies*. Springer, New York (2010)

- Silber, J.H., Rosenbaum, P.R., Ross, R.N., Ludwig, J.M., Wang, W., Niknam, B.A., Mukherjee, N., Saynisch, P.A., Even-Shoshan, O., Kelz, R.R., Fleisher, L.A.: Template matching for auditing hospital cost and quality. *Health Serv. Res.* **49**(5), 1446–1474 (2014a). <https://doi.org/10.1111/1475-6773.12156>
- Silber, J.H., Rosenbaum, P.R., Ross, R.N., Ludwig, J.M., Wang, W., Niknam, B.A., Saynisch, P.A., Even-Shoshan, O., Kelz, R.R., Fleisher, L.A.: A hospital-specific template for benchmarking its cost and quality. *Health Serv. Res.* **49**(5), 1475–1497 (2014b). <https://doi.org/10.1111/1475-6773.12226>
- Stan Development Team. (2021). Stan: A C++ library for probability and sampling (Version 2.26.1) [Computer software]. <http://mc-stan.org/>
- Tang, T., Austin, P.C., Lawson, K.A., Finelli, A., Saarela, O.: Constructing inverse probability weights for institutional comparisons in healthcare. *Stat. Med.* **39**(23), 3156–3172 (2020). <https://doi.org/10.1002/sim.8657>
- Varewyck, M., Goetghebeur, E., Eriksson, M., Vansteelandt, S.: On shrinkage and model extrapolation in the evaluation of clinical center performance. *Biostatistics* **15**(4), 651–664 (2014). <https://doi.org/10.1093/biostatistics/kxu019>
- Zimmerman, L., Olson, H., PMA2020 Principal Investigators Group, Tsui, A., Radloff, S.: PMA2020: rapid turn-around survey data to monitor family planning service and practice in ten countries. *Stud. Fam. Plann.* **48**(3), 293–303 (2017). <https://doi.org/10.1111/sifp.12031>
- Zubizarreta, J.R., Reinke, C.E., Kelz, R.R., Silber, J.H., Rosenbaum, P.R.: Matching for several sparse nominal variables in a case-control study of readmission following surgery. *Am. Stat.* **65**(4), 229–238 (2011). <https://doi.org/10.1198/tas.2011.11072>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.