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

Lucas Godoy Garraza[1], Carolina Cardona[2], Peter Gichangi[3,4,5], Mary Thiongo[3,4,5]

Philip Anglewicz[2], Leontine Alkema[1]

[1] Department of Biostatistics & Epidemiology, University of Massachusetts Amherst
[2] Department of Population, Family and Reproductive Health, Johns Hopkins Bloomberg School of Public Health
[3] Technical University of Mombasa, Kenya
[4] Department of Public Health and Primary Care, Faculty of Medicine and Health Sciences, Ghent University, Belgium
[5] International Centre for Reproductive Health, Mombasa, Kenya

Acknowledgements:

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.

Abstract

Background: 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.

Approach: 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, 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.

Findings: We find significant evidence of variation in FP outcomes across groups of facilities that cannot be accounted for by differences in patient 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.

Table of Contents

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 & 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 woman were not randomly assigned to the facility where they received FP services. Different facilities tend to serve different populations. And a different in populations served can drive differences in outcomes, making naive 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 & Clayton, 2014). Unfortunately, only a limited number of variables can be handled with this model-agnostic approach. Recently, developments from the field of casual inference field have greatly extended the applicability of direct standardization. A notable example is template matching (Silber, Rosenbaum, Ross, Ludwig, Wang, Niknam, Mukherjee, et al., 2014; Silber, Rosenbaum, Ross, Ludwig, Wang, Niknam, Saynisch, et al., 2014), 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 collogues' 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.

Data and measures

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?attachment

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 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.

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 **Error! Reference source not found.**.

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—this procedure is explained in detail in the **Error! Reference source not found.**—yielding a sample of 61 clusters of facilities. In the remainder of the text, facility-level outcomes refer to average outcomes in the clusters.

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 1. Satisfaction and discontinuation are the primary outcomes of interest.

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. Satisfaction is treated both as a proximal outcome and as a potential predictor of discontinuation.

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. The discontinuation rate is the proportion of women who discontinue contraceptive use out of those who received contraceptive method at baseline and have not switched to an alternative method or stopped using a contraceptive method with the intent to become pregnant. In our sample, the outcome was defined for 77% of the women recruited at baseline.

Methods

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^Srefers 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.

Estimating standardized facility-level discontinuation rates

<u>The issue</u>

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 $\mu_j^O = \frac{1}{n_j}\sum_{Z_i=j} m_j^O(X_i)$ 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.

<u>Target quantity</u>

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),

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

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.

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.

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 $\hat{\gamma}_i$:

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

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 page 4. 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.

Especially in smaller clusters, some imbalance may remain after weighting. We cannot directly measure the impact of that residual imbalance on the difference between our estimate, $\hat{\mu}_j^{*W,O}$, and our target, μ_j^*. Nevertheless, we can estimate that difference and remove the estimated bias if we advance a model for the relationship between the outcome and the individual covariates, say $\widehat{m}_j^{BC,O}(x)$. Specifically, given some estimate of the conditional expectation, $\widehat{m}_j^{BC}(x)$, the bias-adjusted estimate is

$$\hat{\mu}_j^{*BC,O} = \hat{\mu}_j^{*W,O} + \left[\frac{1}{n}\sum_{i=1}^{n} \widehat{m}_j^{BC,O}\left(X_i^{tr}\right) - \sum_{i:Z_i=j} \hat{\gamma}_i \widehat{m}_j^{BC,O}\left(X_i^{tr}\right)\right], \tag{5}$$

where the term is brackets is the difference between a simple average of the fitted values from $\widehat{m}_j^{BC}(x)$ over the entire sample and a weighted average of the fitted values over the cluster-specific sample with weights selected so the distribution of the covariates in the cluster resembles the overall distribution. Naturally, this difference is zero if the weights balance the covariance distribution perfectly.

The resulting estimator is termed "bias-corrected", $\hat{\mu}_j^{*BC,O}$. We construct bias-adjusted estimates based on a linear regression model for each of our outcomes of interest, using cluster-specific intercepts and the same set of covariates used for obtaining the weights (see Appendix II, page 8).

Examining outcome variation and the association between facility-level characteristics and standardized outcomes

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 & Pigott, 2001). The Q-statistic is used in meta-analysis to assess heterogeneity across studies. The Q statistics is given by

$$Q^O = \sum_{j}^{n} \frac{\left(\hat{\mu}_j^{*O} - \bar{\mu}^O\right)^2}{\left(\widehat{se}_j^O\right)^2 + (\tau^O)^2} \tag{6}$$

where $\bar{\mu}^{0} = \frac{1}{J}\sum_j \hat{\mu}_j^{*0}$, and $\widehat{se}_j^{O}$ captures estimation error (i.e., the discrepancy between $\hat{\mu}_j^{*0}$ and μ_j^{*O}, see **Error! Reference source not found.**). Cross-cluster variation in outcomes beyond estimation error is captured by τ^{O} . Under the null hypothesis $H_0: \tau^{0} = \tau_0^{0},$ the statistic approximates a χ_{n-1}^2 distribution. While this fact is typically used for hypothesis testing, it can also be used to identify a range of values of τ^{0} with Q values that would not be rejected by the test for a given level (i.e., with p values smaller than, say, α) and among them, the value of τ with less evidence against (i.e., with smallest p value). [1]

As a final step, we assess the association between the standardized facility-level outcomes and facility-level characteristics. 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,

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

$$e_j^S \mid \widehat{se}_j^{BC,S} \sim N\left(0, \left(\widehat{se}_j^{BC,S}\right)^2\right)$$

$$\theta_j^S \mid \tau \sim N(0, (\tau^S)^2)$$

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., $\widehat{se}_j^{BC,S}$) 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. For discontinuation, we therefore pose instead

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

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

[1] The procedures is implemented in the package *blkvar* (Miratrix & Pashley, 2023)

$$\theta_j^B \mid \tau^B \sim N(0, (\tau^B)^2),$$

where $\hat{V}_j^* \equiv n_j^{eff} \times \hat{\mu}_j^{*D,BC}$, 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 in **Error! Reference source not found.**).

Additional analyses

A different approach to compare performance across facilities while accounting for differences in the populations served is to directly model the observed outcomes (without any weights) as a function of both individual- and facility- level characteristics using a hierarchical regression model. The approach is the basis for so called indirect standardization - as opposed to direct standardization we used here. We implemented this alternative approach (see Appendix IV) and compare the facility-level predictors we identify in each case.

We tested the sensitivity of our findings by varying the weights used in the standardization. Specifically, we obtained weights with a different trade-off between improving balance versus improving precision of the estimates (see **Error! Reference source not found.**).

Results

Sample characteristics before and after weighting

Selected women characteristics are included in Table 1. 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.

Variable	**Mean**	**IQR difference**	
		Before weighting	After weighting
Age	28.921	1.344	0.002
Married	0.841	0.091	0.012
Education			
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

Variable	**Mean**	**IQR difference**	
		Before weighting	After weighting
Contraceptive method before this visit			
No method	0.154	0.101	0.003
Same method	0.568	0.149	0.005
Another method	0.279	0.087	0.004
Type of FP given at baseline			
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

Table 1 Women characteristics. Overall average in the sample and interquartile range (IQR) across facilities before and after weighting.

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 and bias-correction. 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.

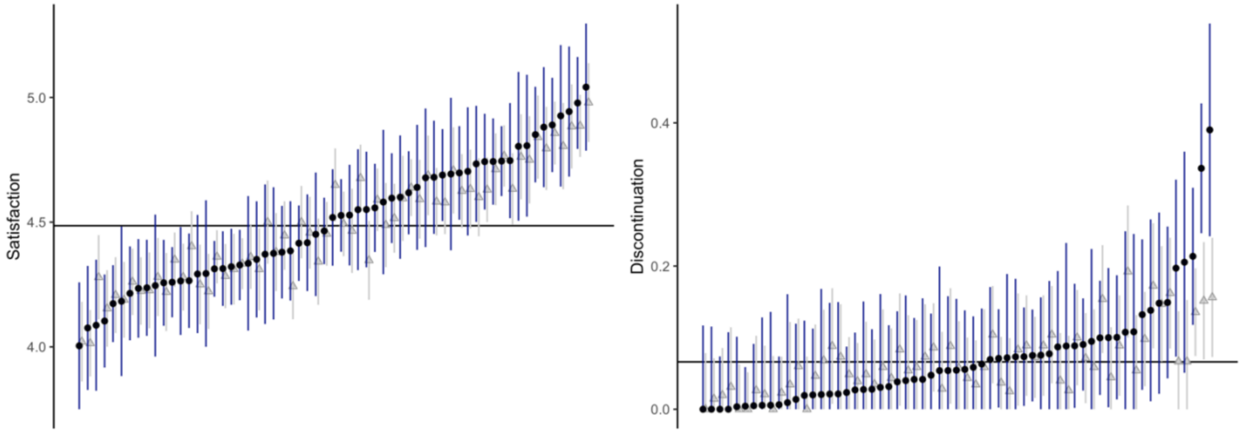

Figure 1 Standardized facility-level discontinuation rate and satisfaction score (blue), obtained after weighting and bias correction; raw rates/scores are included in the background (gray).

Table 2 summarize the results of testing the hypothesis that the variation is entirely due to the uncertainty in the estimates. There is considerable evidence of "true" variation in both satisfaction and discontinuation across facilities that is not accounted by differences in women's characteristics. Specifically, for the bias-corrected estimates, the Q statistics equals 336 and 110 for satisfaction and discontinuation, respectively, unlikely under the null hypothesis of no true variation (p-values < 0.001). The 95% confidence interval for the standard deviation across clusters,τ, is given by (0.190, 0.278) in the case of satisfaction and by (0.032, 0.066) in the case of discontinuation.

Outcome Method	**Grand Average**	**Standard deviation (τ)**		**Q**	**p value**
		Est.	**95% CI**		
Satisfaction					
Unadjusted	4.479	0.211	0.184–0.259	564	0.000
Bias-Corrected	4.481	0.216	0.190–0.278	336	0.000
Discontinuation					
Unadjusted	0.061	0.024	0.013–0.038	89	0.0092
Bias-Corrected	0.064	0.044	0.032–0.066	110	0.0001

Table 2 Average satisfaction score and contraceptive discontinuation rates and estimated standard deviation of the rates and scores across clusters. Confidence Intervals (CI) for the standard deviation are obtained by test inversion, the point estimate is the value with the largest p value.

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 2.5th to 97.5th percentile (see Appendix IV page 14 for further details).

The first three columns of Table 3 show the predictors with APDs with a high probability of being in the direction of the point estimate - above 95% posterior probability using our main approach. 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 or health clinics 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 .224 (SD = .0268, similar to the findings in Table 2) before including any predictors to .193 (SD =.0309) after predictors were included. In the case of discontinuation, we estimated an across-facility standard deviation of .037 (SD = .007) before accounting for predictors. This estimate increased to .051 (SD = .013) after predictors were included, indicating the limited predictive power of the covariate set.

The last three columns of Table 3 include results from a conventional unweighted approach (the model for the conventional approach is described **Error! Reference source not found.**) 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 more positive estimate for provision of postnatal services.

Sensitivity analyses are given in Appendix V and suggest that findings are robust to 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).

Outcome Predictor (percentiles 2.5th, 97.5th)	Balancing weights with λ = .001			Balancing weights with λ = .1			Unweighted model-based adjustment		
	APD	SD	P	APD	SD	P	APD	SD	P
Satisfaction									
Facility type: dispensary (0, 1)	0.199	0.098	0.98	0.190	0.086	0.986	0.202	0.086	0.991
Staff here/ total staff (0.29, 0.9)	0.264	0.134	0.972	0.279	0.118	0.991	0.277	0.113	0.991
FP visits last month (90.2, 643.87)	−0.280	0.148	0.969	-0.320	0.130	0.990	−0.320	0.129	0.993
Facility is public (0.66, 1)	−0.507	0.279	0.962	-0.534	0.252	0.984	−0.540	0.251	0.986
Facility type: Health Clinic (0, .1)	−0.548	0.338	0.951	-0.471	0.301	0.942	−0.356	0.300	0.888
Facility type: hospital (0, 1)	0.169	0.150	0.876	0.209	0.128	0.950	0.267	0.131	0.979
Discontinuity									
Staff to visits ratio (0.01, 0.63)	−0.099	0.060	0.955	−0.075	0.043	0.957	0.080	0.106	0.781
Postnatal services (0.6, 1)	0.067	0.060	0.945	0.007	0.106	0.712	0.061	0.023	0.964

Table 3 Comparison of average predictive difference (APD) between direct standardization and hierarchical model for selected predictors. 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. The "P" value is the posterior probability of APD having the sign of its point estimate.

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 use newly collected data from the PMA project that allows 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 allows 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 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.

Our findings were similar across different approaches using standardization but differed from a more conventional regression model approach. 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 facility level characteristics and outcomes of interest are not necessarily causal ones. Causal interpretation of the analysis would require additional assumptions, such as randomness in the distribution of the specific facility covariates.

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.

References

Ali, M. M., & Cleland, J. (2010). Contraceptive Switching after Method-related Discontinuation: Levels and Differentials. *Studies in Family Planning*, *41*(2), 129–133. https://doi.org/10.1111/j.1728-4465.2010.00234.x

Anglewicz, P., Cardona, C., Akinlose, T., Gichangi, P., OlaOlorun, F., Omoluabi, E., Thiongo, M., Akilimali, P., Tsui, A., Kayembe, P., & Group, T. P. A. P. I. (2021). Service delivery point and individual characteristics associated with the adoption of modern contraceptive: A multi-country longitudinal analysis. *PLOS ONE*, *16*(8), e0254775. https://doi.org/10.1371/journal.pone.0254775

AssunÇão, R. M., Neves, M. C., Câmara, G., & Da Costa Freitas, C. (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, *20*(7), 797–811. https://doi.org/10.1080/13658810600665111

Bellow, N., Dougherty, L., Nai, D., Kassegne, S., Nagbe, R. H. Y., Babogou, L., Guede, K. M., & Silva, M. (2023). Improving provider and client communication around family planning in Togo: Results from a cross-sectional survey. *PLOS Global Public Health*, *3*(6), e0001923. https://doi.org/10.1371/journal.pgph.0001923

Bloom, H. S., Raudenbush, S. W., Weiss, M. J., & Porter, K. (2017). Using Multisite Experiments to Study Cross-Site Variation in Treatment Effects: A Hybrid Approach With Fixed Intercepts and a Random Treatment Coefficient. *Journal of Research on Educational Effectiveness*, *10*(4), 817–842. https://doi.org/10.1080/19345747.2016.1264518

Bradley, S. E., Schwandt, H. M., & Khan, S. M. (2009). *Levels, Trends, and Reasons for Contraceptive Discontinuation*.

Breidt, F. J., & Opsomer, J. D. (2017). Model-Assisted Survey Estimation with Modern Prediction Techniques. *Statistical Science*, *32*(2). https://doi.org/10.1214/16-STS589

Bürkner, P.-C. (2017). **brms**: An *R* Package for Bayesian Multilevel Models Using *Stan*. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Cardona, C., OlaOlorun, F. M., Omulabi, E., Gichangi, P., Thiogo, M., Tsui, A., & Anglewicz, P. (2022). The relationship between client dissatisfaction and contraceptive discontinuation among urban family planning clients in three sub-Saharan African countries. *PLOS ONE*, *17*(8), e0271911. 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. (2019). Association Between the Quality of Contraceptive Counseling and Method Continuation: Findings From a Prospective Cohort Study in

Social Franchise Clinics in Pakistan and Uganda. *Global Health: Science and Practice*, *7*(1), 87–102. https://doi.org/10.9745/GHSP-D-18-00407

Chen, C., Wakefield, J., & Lumely, T. (2014). The use of sampling weights in Bayesian hierarchical models for small area estimation. *Spatial and Spatio-Temporal Epidemiology*, *11*, 33–43. https://doi.org/10.1016/j.sste.2014.07.002

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, *19*(1). https://doi.org/10.1214/aos/1176347963

Gagolewski, M., Cena, A., Bartoszuk, M., & Brzozowski, Ł. (2023). *Clustering with minimum spanning trees: How good can it be?* (arXiv:2303.05679). arXiv. http://arxiv.org/abs/2303.05679

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. https://doi.org/10.1214/06-BA117A

Gelman, A., & Pardoe, I. (2007). 2. Average Predictive Comparisons for Models with Nonlinearity, Interactions, and Variance Components. *Sociological Methodology*, *37*(1), 23–51. https://doi.org/10.1111/j.1467-9531.2007.00181.x

Hartung, J., Knapp, G., & Sinha, B. K. (2008). *Statistical Meta-Analysis with Applications*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470386347

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, *6*(3), 203–217. https://doi.org/10.1037/1082-989X.6.3.203

Jain, A., Aruldas, K., Mozumdar, A., Tobey, E., & Acharya, R. (2019). Validation of Two Quality of Care Measures: Results from a Longitudinal Study of Reversible Contraceptive Users in India. *Studies in Family Planning*, *50*(2), 179–193. https://doi.org/10.1111/sifp.12093

Karp, C., OlaOlorun, F. M., Guiella, G., Gichangi, P., Choi, Y., Anglewicz, P., & Holt, K. (2023). 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. *Studies in family planning*, *54*(1), 119-143.

Keele, L., Ben-Michael, E., Feller, A., Kelz, R., & Miratrix, L. (2021). *Hospital Quality Risk Standardization via Approximate Balancing Weights* (arXiv:2007.09056). arXiv. http://arxiv.org/abs/2007.09056

Keiding, N., & Clayton, D. (2014). Standardization and Control for Confounding in Observational Studies: A Historical Perspective. *Statistical Science*, *29*(4). https://doi.org/10.1214/13-STS453

Miratrix, L., & Pashley, N. (2023). *blkvar: ATE and Treatment Variation Estimation for Blocked and Multisite RCTs* (0.0.1.5) [Computer software].

Normand, S.-L. T., Ash, A. S., Fienberg, S. E., Stukel, T. A., Utts, J., & Louis, T. A. (2016). League Tables for Hospital Comparisons. *Annual Review of Statistics and Its Application*, *3*(1), 21–50. https://doi.org/10.1146/annurev-statistics-022513-115617

Normand, S.-L. T., Glickman, M. E., & Gatsonis, C. A. (1997). Statistical Methods for Profiling Providers of Medical Care: Issues and Applications. *Journal of the American Statistical Association*, *92*(439), 803–814. https://doi.org/10.1080/01621459.1997.10474036

Oyugi, B., Kioko, U., Kaboro, S. M., Okumu, C., Ogola-Munene, S., Kalsi, S., Thiani, S., Gikonyo, S., Korir, J., Baltazar, B., & Ranji, M. (2018). A facility-based study of women' satisfaction and perceived quality of reproductive and maternal health services in the Kenya output-based approach voucher program. *BMC Pregnancy and Childbirth*, *18*(1), 310. https://doi.org/10.1186/s12884-018-1940-9

Potthoff, R. F., Woodbury, M. A., & Manton, K. G. (1992). "Equivalent Sample Size" and "Equivalent Degrees of Freedom" Refinements for Inference Using Survey Weights under Superpopulation Models. *Journal of the American Statistical Association*, *87*(418), 383–396. https://doi.org/10.1080/01621459.1992.10475218

Prim, R. C. (1957). Shortest Connection Networks And Some Generalizations. *Bell System Technical Journal*, *36*(6), 1389–1401. https://doi.org/10.1002/j.1538-7305.1957.tb01515.x

Raghunathan, T. E., Xie, D., Schenker, N., Parsons, V. L., Davis, W. W., Dodd, K. W., & Feuer, E. J. (2007). Combining Information From Two Surveys to Estimate County-Level Prevalence Rates of Cancer Risk Factors and Screening. *Journal of the American Statistical Association*, *102*(478), 474–486. https://doi.org/10.1198/016214506000001293

Rosenbaum, P. R. (2010). *Design of observational studies*. Springer.

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. (2014). Template Matching for Auditing Hospital Cost and Quality. *Health Services Research*, *49*(5), 1446–1474. 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. (2014). A Hospital-Specific Template for Benchmarking its Cost and Quality. *Health Services Research*, *49*(5), 1475–1497. https://doi.org/10.1111/1475-6773.12226

Stan Development Team. (2021). *Stan: A C++ Library for Probability and Sampling* (2.26.1) [Computer software]. http://mc-stan.org/

Tang, T., Austin, P. C., Lawson, K. A., Finelli, A., & Saarela, O. (2020). Constructing inverse probability weights for institutional comparisons in healthcare. *Statistics in Medicine*, *39*(23), 3156–3172. https://doi.org/10.1002/sim.8657

Zimmerman, L., Olson, H., PMA2020 Principal Investigators Group, Tsui, A., & Radloff, S. (2017). PMA2020: Rapid Turn-Around Survey Data to Monitor Family Planning Service and Practice in Ten Countries. *Studies in Family Planning*, *48*(3), 293–303. https://doi.org/10.1111/sifp.12031

Zubizarreta, J. R., Reinke, C. E., Kelz, R. R., Silber, J. H., & Rosenbaum, P. R. (2011). Matching for Several Sparse Nominal Variables in a Case-Control Study of Readmission Following Surgery. *The American Statistician*, *65*(4), 229–238. https://doi.org/10.1198/tas.2011.11072