Evaluation of Large-Scale Maternal and Child Health Programs in India & Kenya

Evaluation of Large-scale Maternal and Child Health Programs in India & Kenya

Dale A. Barnhart

A Dissertation Submitted to the Faculty of

The Harvard T.H. Chan School of Public Health

in Partial Fulfillment of the Requirements

for the Degree of Doctor of Science

in the Department of Epidemiology

Harvard University

Boston, Massachusetts.

May 2019

Dissertation Advisor: Donna Spiegelman Dale A. Barnhart

Evaluation of Large-Scale Maternal and Child Health Programs in India & Kenya

Abstract

Major reductions in morbidity and mortality could be achieved through high-quality, high-coverage implementation of proven maternal and child health interventions, including facility-based delivery by skilled birth attendants and prevention of mother-to-child HIV transmission through the provision of antiretroviral drugs. Methodologic advances in implementation science could identify opportunities to improve the effectiveness of these programs. This dissertation illustrates how epidemiologic methods can be applied to implementation science research using data from two large-scale maternal and child health programs; namely the BetterBirth Program, which implemented the World Health Organization's Safe Childbirth Checklist in Uttar Pradesh, India, and the U.S. President's Emergency Plan for AIDS Relief (PEPFAR) funding of prevention of mother-to-child HIV transmission (PMTCT) activities in Kenya. This dissertation A) reviews the development of the BetterBirth Program's implementation package to identify methodologic lessons learned that can assist future researchers seeking to develop and evaluate complex interventions; B) performs regression analyses to investigate the relationship between coaching intensity, adherence to essential birth practices (EBPs), and maternal and perinatal health outcomes in the BetterBirth study in order to provide insights into the optimal coaching regimen for future interventions; and C) estimates the impact of per capita PEPFAR funding for PMTCT on infant and neonatal mortality and HIV counseling, testing, and receipt of test results during antenatal care in Kenya using a

quasi-experimental dose-response analysis of publicly available data. These evaluations illustrate how the collection and analysis of quantitative data can advance our understanding of the overall effectiveness of large-scale maternal and child health programs and provide insights into how the implementation of these programs should be improved.

Table of Contents

List of Figures with Captions

List of Tables with Captions

Acknowledgments

This dissertation would not have been possible without the support of my advisor, Donna Spiegelman. I am grateful for her willingness to take on such an inexperienced student and especially thankful for her generosity with her time over these years.

I would additionally like to thank Katherine Semrau for her ongoing professional and personal support and openness to collaboration as well as Cory Zigler for consistently responding to my questions and concerns with helpful insights. I would also like to acknowledge the contributions of the larger BetterBirth team as well as my colleagues at the Henry M. Jackson Foundation Medical Research International, Kericho, Kenya; the U.S. Military HIV Research Program, the Office of the U.S. Global AIDS Coordinator and Health Diplomacy and the Henry M. Jackson Foundation for the Advancement of Military Medicine.

To my friends in Boston, most especially Haley, Sam, Sarah, and Hailley, I would not have finished this without your daily support. To my family, particularly Mom and Dad, thank you for encouraging me to pursue not just this opportunity but also all the other opportunities that made this possible. I would not have gotten here without you.

Chapter One: Optimizing the development and evaluation of complex interventions: Lessons learned from the BetterBirth Program and associated trial

ABSTRACT:

Background: Despite extensive efforts to develop and refine optimized intervention packages, complex interventions often fail to produce the desired health impacts in full-scale evaluations. A recent example of this phenomenon is BetterBirth, a complex intervention designed to implement the World Health Organization's Safe Childbirth Checklist and improve maternal and neonatal health. Using data from the BetterBirth Program and its associated trial as a case study, we identified lessons to assist in the development and evaluation of future complex interventions.

Methods: BetterBirth was refined across three sequential development phases prior to being tested in a matched-pair, cluster-randomized trial in Uttar Pradesh, India. We reviewed published and internal materials from all three development phases to identify barriers hindering the identification of an optimal intervention package and identified corresponding lessons learned. For each lesson, we describe its importance and provide an example inspired by the BetterBirth Program's development to illustrate how it could be applied to future studies.

Results: We identified three lessons: 1) Develop a robust theory of change (TOC); 2) Define optimization outcomes, which are used to compare the effectiveness of the intervention across development phases, and corresponding criteria for success, which are used to determine whether the intervention has been sufficiently optimized to warrant full-scale evaluation; and

3) Create and capture variation in the implementation intensity of components. When applying these lessons to the BetterBirth intervention, we demonstrate how a TOC could have promoted more complete data collection. We propose an optimization outcome and related criteria for success and illustrate how they could have resulted in additional development phases prior to the full-scale trial. Finally, we show how variation in components' implementation intensities could have been used to identify effective intervention components.

Conclusion: These lessons learned can be applied during both early and advanced stages of complex intervention development and evaluation. By using examples from a real-world scenario to demonstrate the relevance of these lessons and illustrating how they can be applied in practice, we hope to encourage future researchers to collect and analyze data in a way that promotes more effective development and evaluation of complex interventions.

BACKGROUND

Complex interventions consist of a package of several interacting components (1, 2) and are widely used in public health applications including HIV prevention (3), smoking cessation (4), and prevention of childhood obesity (5). Complex interventions are ideally both effective, or able to produce a heath impact, and optimized, or able to efficiently use available resources in a way that produces the greatest possible heath impact. Currently, there is little consensus on how to best develop complex interventions. Methodologically rigorous approaches, such as the factorial or fractional-factorial designs often used in the Multiphase Optimization Strategy (MOST) framework, can estimate causal effects of individual package components (6, 7). However, these designs require researchers to specify detailed information on candidate components at the beginning of the study, which may not be feasible early in the development process. They can also be prohibitive in cluster-randomized studies where few units are available for randomization (8) or the cost per unique treatment condition is high (9). The recently developed Learn-as-You-Go (LAGO) design allows researchers to estimate the effects of individual components using data collected in phases, with data from previous phases being used to recommend interventions for subsequent phases (10). However, this design has yet to be used in a real-world study.

In practice, complex interventions are often developed and refined using expert and stakeholder consensus and qualitative research (11-14). These approaches are rarely accompanied by quantitative analyses illustrating the effectiveness of individual implementation components or demonstrating that the intervention has been sufficiently optimized to warrant a full-scale evaluation. Consequently, many interventions fail to produce

the desired impact on health outcomes in a full-scale trial (15). One recent example of this phenomenon is an intervention called BetterBirth, a complex intervention designed to improve the quality of care in childbirth facilities with the ultimate goal of improving maternal and neonatal health. Despite extensive preliminary research during the intervention development process (16), the intervention did not improve maternal and child health in a recent high-profile trial, although it did improve birth attendant adherence to evidence based practices (17). This paper identifies lessons learned from the BetterBirth experience and provides illustrative examples showing how these lessons could be applied to the development and evaluation of future complex interventions.

METHODS

Overview of the BetterBirth intervention

BetterBirth is a complex intervention consisting of a multi-component implementation package designed to promote the use of the World Health Organization's Safe Childbirth Checklist (SCC). The 28-item SCC is intended to help birth attendants successfully complete evidence-based essential birth practices (EBPs) that prevent or successfully manage complications during facility-based deliveries (18). BetterBirth was developed through a multi-phase process. The initial implementation package was informed by team members' prior experiences successfully implementing a similar quality of care improvement tool, the Safe Surgical Checklist (19, 20) and by a pilot study of the SCC conducted in a hospital in Karnataka, India (21). This initial package was refined over three sequential development phases conducted in primary-level health facilities in Uttar Pradesh, India. The first two phases, Pilot 1 and Pilot 2, were pre-post studies

conducted in two and four facilities, respectively. The third development phase occurred among the first 15 control-intervention pairs enrolled in a matched-pair, cluster-randomized trial (CRT) designed to assess effectiveness of the BetterBirth intervention on reducing maternal morbidity and maternal and infant mortality (22). We consider these 15 pairs to constitute a development phase because researchers originally planned to conduct preliminary analyses among these facilities and further adapt the intervention as needed prior to enrolling the remaining CRT facilities. However, time and budgetary constraints ultimately prevented further adaptations to the final implementation package. To accommodate both the pre-post and matched-pair, cluster-randomized designs, we designate all births occurring in a control site or any births occurring in an intervention site prior to the introduction of the BetterBirth intervention as part of the "control period" and any birth occurring in an intervention site after the introduction of the BetterBirth intervention as part of the "intervention period."

Across the three development phases, the content, delivery, and intensity of implementation package components assigned to facilities varied, as described in Hirschhorn et al. (16) and summarized in Table 1.1. During Pilot 1, the BetterBirth intervention included 3 package components: leadership engagement, an educational and motivational program launch, and ongoing coaching visits to promote checklist use and EBP adherence. The fourth package component, a data feedback cycle in which birth attendants were provided with quantitative information on their performance, was added to the Pilot 2 and CRT phases. In each phase, the intervention's effectiveness was assessed based on birth attendants' SCC use and EBP adherence, which was directly observed by trained independent nurses and recorded using standardized data collection tools. EBP observations occurred during three distinct periods of

delivery: on admission to facility, just before pushing, and within one hour after birth. However, practical considerations related to the timing and duration of labor prevented all births from being continuously observed from admission through discharge such that not all EBPs were observed for each birth. Data on EBP adherence was available on 113 births from Pilot 1; 2,369 births from Pilot 2; and 6,562 births from the CRT phase.

Table 1.1 The BetterBirth implementation package by phase

Phase	Leadership engagement	Educational & Motivational Launch	Data Feedback	Coaching visits
Pilot Phase 1	Non-standard initial engagement, with a focus on facility rather than district leadership	3-day launch featuring one day of flipchart & video-based training, one day of checklist demonstrations and placement of checklist posters on walls, and one day of facilitated practice sessions on checklist use.	None	1 coaching visit every 2 weeks for the first 6 months, then 1 coaching visit per month.
Pilot Phase 2	Standardized initial engagement with district and facility leadership	Semi-standardized 2-day launch featuring flipchart; videos; checklist posters; roleplaying; and the identification of a Childbirth Quality Coordinator	Ongoing feedback, using paper-based reports. Frequency of report generation and delivery to sites unspecified.	3 coaching visits per week for the first 4 weeks, then less frequently.
CRT	Standardized initial engagement with district and facility leadership. Semi-regular meetings with district leadership	Standardized 2-day launch featuring flipchart; videos; checklist posters; roleplaying; the identification of a Childbirth Quality Coordinator; and a safe-childbirth pledge.	Ongoing feedback using app-based reports. Frequency of report generation and frequency of sites reviewing feedback in the app are unspecified.	2 coaching visits per week for months 1-4; 1 coaching visit per week for months 5-6; 1 coaching visit per fortnight in month 7; 1 coaching visit per month in month 8.

Identifying lessons learned

To identify barriers preventing the identification an optimal BetterBirth implementation package, we reviewed published articles, research protocols, internal reports, data collection tools, implementation team weekly updates, and data from all three development phases of the BetterBirth intervention. The results of this review were used to identify both barriers that hindered the identification of an optimal intervention package and corresponding lessons learned. For each lesson, we described its importance and used material inspired by the BetterBirth Program to illustrate how this lesson could be applied in practice. These illustrative examples were designed to aid in the development and evaluation of future complex interventions.

Illustrative examples and analysis

Our theory of change (TOC) was drafted after our initial review and refined following a group discussion with members from the BetterBirth team. To assess the overall effectiveness of the intervention, we used a generalized linear model with adjusted for the development phases (Pilot 1, Pilot 2, and CRT phases), the intervention (vs. control) period, and their interactions (23). When assessing the coaching component, we additionally added coaching intensity, calculated for each birth as the number of coaching visits that occurred at their facility in the 30 days prior to their birth (24). In Pilot 2, only the first and last dates of coaching and the total number of coaching visits per site were recorded. In order to calculate coaching intensity metrics, we imputed these missing coaching dates under the assumption that they followed a uniform distribution bounded by the first and last dates of coaching. All models accounted for

clustering at the facility level by estimating standard errors using the empirical variance with an exchangeable working covariance structure.

RESULTS

We identified three key lessons learned: 1) Develop a robust theory of change, 2) Define optimization outcomes, which are used to compare the effectiveness of the intervention across development phases, and corresponding criteria for success, which are used to determine whether the intervention has been sufficiently optimized to warrant full-scale evaluation, and 3) Create and capture variation in the implementation intensity of intervention components. For each lesson, we describe its importance, discuss how it applies to the BetterBirth Program, and provide an illustrative example.

Lesson 1: Develop a robust theory of change

Theories of change (TOC) are tools that define and communicate researchers' underlying assumptions and hypotheses about the processes through which a complex intervention improves outcomes (25-27). The assumptions and hypotheses encoded in a TOC can be informed by a wide range of generalized theories commonly used in implementation science (28) including classical theories from psychology and sociology like the Theory of Planned Behavior (29) and implementation theories like the Theoretical Domains Framework (30). However, TOCs differ from generalized theories because they describe causal relationships between variables in a way that is specific to both the intervention of interest and the context in which that intervention is being implemented (26, 27). TOCs should include the complex intervention's individual components, primary outcome, and any process outcomes

hypothesized to be on the causal pathway between at least one intervention component and the primary outcome. Additionally, TOCs should contain information on contextual factors expected to impact the relationship between these variables. Many researchers use the terms logic model and TOC interchangeably, and some logic models include sufficiently detailed causal pathways such that they can function as a TOC (e.g. 31). However, while TOCs necessarily include information about the assumed causal connections between variables, logic models often assume simplistic progressions between groups of variables, such as inputs, outputs, outcomes, and impacts (e.g. 32, 33) without making their causal assumptions explicit (26, 27).

The causal assumptions contained in a TOC provide a structure for identifying and addressing the challenging hallmarks of complex intervention research (1, 2). By identifying which hypothesized causal links are thought to be of greatest importance to the overall success of the intervention (or, alternatively, are the subject of greatest uncertainty), TOCs help prioritize data collection and guide subsequent data analyses (2, 25, 26, 34). Testing for the existence of the hypothesized causal links identified in the TOC can help identify ineffective intervention components, highlight incorrect assumptions about the underlying mechanism of change or context in which the intervention is being implemented, and inform future adaptions to the intervention (26, 34-36). TOCs can also identify the appropriate data source and unit of analysis for each variable and can highlight which data sources will need to be linked together for analysis (37, 38). Finally, TOCs can strengthen collaborations between interdisciplinary team members who may not otherwise share common assumptions or vocabulary for describing the intervention (25, 26, 39).

Lesson 1: Application to BetterBirth

The BetterBirth team used two theoretical frameworks to develop their intervention: the "Engage, Educate, Execute, and Evaluate" framework in Pilot 1 (Figure 1.1a) and the "Engage, Launch, Support" framework in the Pilot 2 and CRT phases (Figure 1.1b). Although these frameworks were grounded in a generalized model known as the "4-Es" (40), they did not constitute a TOC because as they did not contain information on the specific causal pathways through which individual implementation package components are hypothesized to improve maternal and child health. While these frameworks were effective at communicating the program's overall implementation strategy, they could not be used to prioritize data collection, guide analyses, or suggest opportunities for adapting and improving the intervention. In parallel with these frameworks, an informal TOC developed among BetterBirth team members (Figure 1.1c). However, this TOC was too simplified to inform data collection or analysis and, without formalized language and broad stakeholder buy-in, could not strengthen collaborations among team members (41).

We produced an alternative, robust TOC for the BetterBirth intervention (Figure 1.2). We represented this TOC using arrows to designate hypothesized causal relationships, shading to identify the desired unit of analysis for each variable, and superscripts to identify which variables were measured in each BetterBirth development phase. Many variables that were identified in this TOC as playing an important role in the BetterBirth Program, such as birth attendant ability, were not measured during the development phases. Other variables, such as birth attendant attitudes towards the SCC, were measured in only one phase and therefore could not be compared across phases.

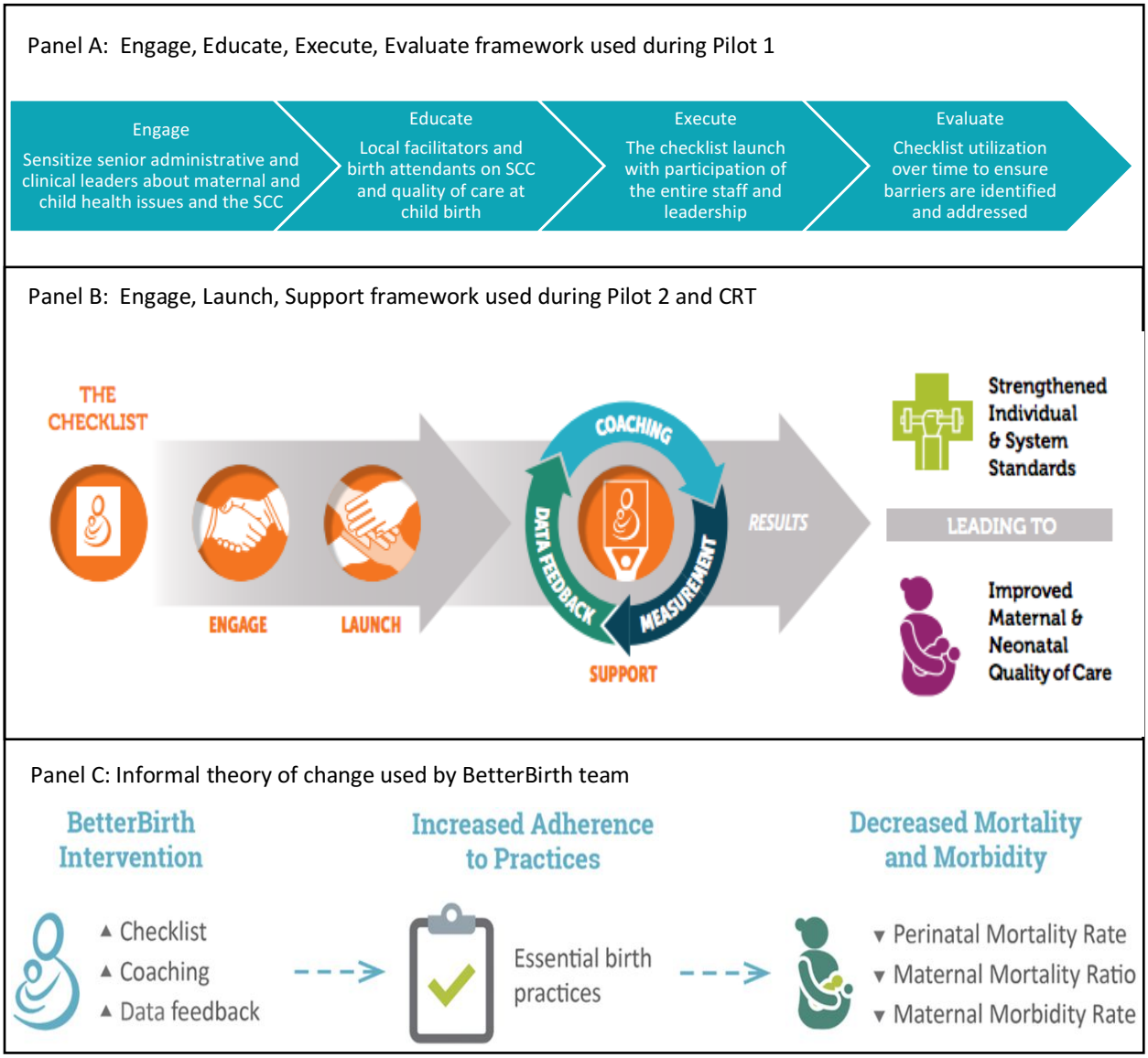

Figure 1.1 Frameworks used during the development of the BetterBirth intervention

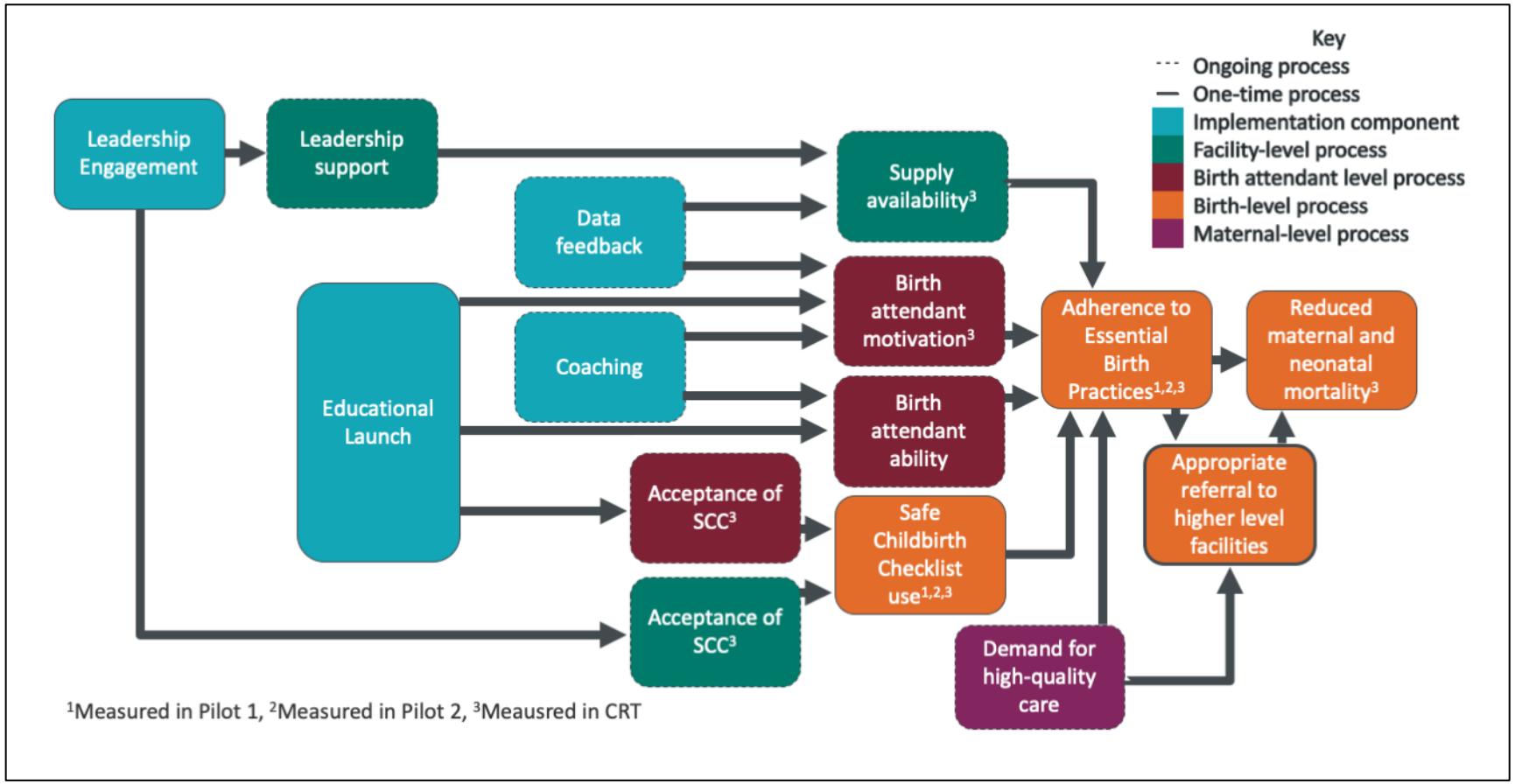

Figure 1.2 Robust theory of change for the BetterBirth intervention

This practice prevented deep understanding on the impact of changes made to the implementation package on the intervention. Finally, some hypothesized casual links in the TOC existed between variables that were assessed at different units of analysis. For example, attitude towards the SCC, which was assessed at the birth attendant level, was hypothesized to impact SCC use, which was assessed at the individual birth level. However, because the data collection process did not allow for individual birth attendants to be linked to individual births, this hypothesized link could only be assessed using data aggregated at the facility level. Developing a robust TOC at the start of the intervention development process could have highlighted these limitations earlier, promoted more complete data collection, and provided additional opportunities to learn about the intervention.

Lesson 2: Select optimization outcomes and specify criteria for success

After creating a TOC, a subset of process outcomes can be selected as optimization outcomes. Optimization outcomes serve two functions during the complex intervention development process. First, comparing optimization outcomes across development phases identifies whether changes to the intervention have improved the intervention's overall effectiveness. Second, comparing optimization outcomes against criteria for success identifies whether the complex intervention has been sufficiently optimized to warrant evaluation in a full-scale trial (42). In the context of intervention development, criteria for success (42), which have also been described as optimization criteria (6) and "Go/No Go" rules (43), describe the minimal effect the intervention should have on the optimization outcomes such that there is a reasonable chance that the intervention could meaningfully impact the primary outcome. Interventions

that fail to meet these criteria require additional phases of development before progressing to a full-scale evaluation.

To serve these two functions, optimization outcomes should be defined and assessed consistently across all development phases. They should also be valid surrogates, or proxies, for the primary outcome. Surrogate outcomes are widely used in clinical efficacy trials when collecting data on the primary outcome is expensive, time-consuming, or otherwise infeasible (43, 44). The effect of the intervention on a valid surrogate outcome will correspond to the effect of intervention on the primary outcome. However, it can be difficult to identify valid surrogate outcomes (45, 46). Several high-profile trials demonstrate how poorly chosen surrogates can lead to misleading conclusions about an intervention's effectiveness (47-51). Multiple strategies for empirically validating surrogate outcomes have been proposed, including a meta-analytic approach which can be used in settings where researchers have access to data on intervention status, optimization outcomes, and primary outcomes from multiple trials or for several sub-populations within a single trial (52, 53). However, this sort of data is usually not available to researchers seeking to developing a new intervention. In the absence of empirical verification, researchers must use their knowledge about the intervention and its expected effects to determine whether a candidate optimization outcome is likely to be a valid surrogate for the primary outcome.

In settings where the intervention is hypothesized to improve the primary outcome, researchers typically select an optimization outcome that is a) also expected to improve as a result of the intervention and b) is positively correlated with the primary outcome. In this setting, the following conditions will nearly always guarantee that the optimization outcome is

a valid surrogate. More general conditions can be found elsewhere (45, 46). First, the positive correlation between the surrogate and the primary outcome should reflect a positive causal effect and not be induced by bias. For example, in BetterBirth, it was believed that increased checklist use would be positively correlated with maternal and child survival. However, checklist use would not have been a valid surrogate for survival if this correlation was explained by confounding rather than causation, as would have occurred if educated birth attendants were both more likely to use the checklist and more likely to have good patient outcomes. Second, if there are mechanisms through which the intervention could have unintended adverse effects on the primary outcome, those mechanisms should also adversely impact the surrogate outcome. For example, it is plausible that the BetterBirth intervention could have increased birth attendants' adherence to specific tasks, such as providing oxytocin immediately after delivery, by decreasing the amount of time and resources they dedicated to other tasks, such as taking the temperature of newborns. In this context, adherence to any single task would have been unlikely to serve as a valid surrogate because it would not have reflected the potential unintended consequences of decreased adherence to other tasks. Third, if improvements in the surrogate are only beneficial among a specific subgroup of individuals, as would be the case in the presence of effect modification, then the intervention should improve the surrogate outcome within that subgroup. For example, in the context of the BetterBirth intervention, antibiotics were only expected to improve survival among the subgroup of mothers and infants who were at-risk of infection. Therefore, increases in antibiotic prescription rates would have only been a valid surrogate for survival if those increases were occurring specifically among mothers and infants identified as being at risk of infection.

In addition to selecting an optimization outcome, researchers must also specify criteria for success. Criteria for success should reflect both the strength of the relationship between the optimization outcome and the primary health outcomes and whether that relationship is likely to exhibit non-linear effects. Situational considerations should dictate whether criteria for success are set in relative or absolute terms. For example, if researchers believe that the primary outcome will only change after the optimization outcome crosses a certain threshold, then criteria for success should be defined with respect to that absolute threshold, not in terms of relative improvements. If multiple process outcomes are selected as optimization outcomes, the criteria for success should account for each of these outcomes, either by combining them into a single composite outcome (e.g. average EBP adherence must reach 86%) or by creating individual success criterion for each outcome (e.g. adherence to each EBP much reach 70%). Selecting optimization outcomes and specifying criteria for success are both critical steps in the optimization process. If either is misspecified, then researchers could develop an intervention that improves the optimization outcome and meets the criteria for success but still fails to impact the primary outcome.

Finally, the sample size for each development phase should be calculated with respect to the optimization outcomes and their corresponding criteria for success. Phases do not necessarily need to be powered for formal hypothesis tests with a 5% Type I error rate (43). However, each phase should be powered such that estimates for the effect of the intervention on the optimization outcome are precise enough to inform a decision to proceed to the full-scale trial. For example, if the criteria for success are defined as observing a confidence interval for the optimization outcome that includes or exceeds some pre-specified value (54), then power

calculations seek to ensure that the confidence intervals for the optimization outcome will be informatively narrow (42, 55).

Lesson 2: Application to BetterBirth

During the BetterBirth intervention's development phases, EBP adherence was used to assess the intervention's effectiveness since the primary health outcomes of maternal morbidity and maternal and infant mortality were relatively rare. During these phases, each EBP was analyzed and reported on independently, with different sets of EBPs being used in different pilot phases (16, 56). Criteria for success were not specified for any EBP. This approach had several limitations. First, it is unlikely that adherence to individual EBPs served as valid surrogates. As previously discussed, adherence to individual tasks may not capture other unintended adverse consequences of the intervention. Furthermore, secondary analysis of the BetterBirth Trial data has suggested that individual EBPs were generally not correlated with improved health outcomes (57). Second, without predefined criteria for success, it is unclear how EBP adherence data informed the decision to progress to a full-scale trial. Although the BetterBirth Trial observed large, statistically significant relative gains in EBP adherence to EBPs remained low. For example, although intervention sites were 53 times more likely to use appropriate hand hygiene than control sites, they still only used appropriate hand hygiene 35% of the time (17). If criteria for success had been defined in terms of absolute EBP adherence, these improvements may have been recognized as too modest to result in meaningful health improvements, which would have triggered additional development phases. Third, these limitations were compounded by inconsistent data collection across the three development phases. Not all EBPs were assessed in all phases, and the timing and duration of EBP data collection relative to the

start of coaching differed from phase to phase. Consequently, observed changes in EBP adherence could have been caused by either real changes in the program's effectiveness or by inconsistencies in data collection. Finally, while the CRT (N=6,562) was powered to detect 8.5 percentage point differences in EBP adherence between intervention and control sites, Pilot 1 (N=113) and Pilot 2 (N= 2,369) had relatively small sample sizes and were underpowered to detect meaningful differences in EBP adherence.

Table 1.2 illustrates how a composite outcome of overall EBP adherence could have served as the optimization outcome and provided additional information about the intervention's effectiveness prior to the full-scale trial. We defined overall EBP adherence as the proportion of observed EBPs that were successfully completed at each birth out of a set of eight EBPs that were measured consistently across all three phases (Table 1.2). This composite outcome was expected to serve as a valid surrogate for maternal and infant survival for several reasons. First, overall EBP adherence was correlated with improved infant survival (57), and it was assumed that this correlation reflected a causal effect. Second, the set of EBPs included in the outcome was include EBPs measured at all three pausepoints and EBPs performed on both the mother and the baby, so if the intervention had produced any unintended adverse effects, these negative effects would have likely resulted in reduced adherence to at least one of the included EBPs. Finally, each EBP included in the composite outcome was believed to be beneficial for all births, not just among a certain subgroup, so increased overall EBP adherence was expected to benefit the entire patient population. The use of overall EBP adherence as an optimization outcome was also supported by our TOC because a) it was proximal to the primary outcome of infant and maternal mortality and b) there were no hypothesized causal pathways from the

intervention components to infant and maternal mortality that did not go through EBP adherence. We defined our criteria for success as observing an intervention period in which the 95% confidence interval for overall EBP adherence included 86%. This threshold was based on a previously successful implementation of the SCC in a hospital in Karnataka, India, where EBP adherence increased from 34% to 86% and researchers observed a marginally significant halving of stillbirths (21). Our analysis suggested that, although the intervention increased EBP adherence in all phases, changes to the implementation package across phases did not fully improve total EBP completion. Furthermore, even though Pilot 1 and Pilot 2 are underpowered and had correspondingly wide confidence intervals, the 95% confidence intervals for overall EBP adherence during the intervention period did not include 86% in any phase, which suggested that the implementation package was not sufficiently optimized at the time of the trial

Table 1.2 Effectiveness of each phase of the BetterBirth intervention on overall Essential Birth Practice (EBP) Adherence, which was calculated as the percentage of observed EBPs that were successfully completed out of eight EBPs: 1) use of a partograph, 2) maternal blood pressure at admission, 3) maternal temperature at admission, 4) appropriate hand hygiene prior to a push, 5) provision of oxytocin to the mother within one minute of delivery, 6) assessment of baby weight, 7) assessment of newborn temperature, and 8) initiation of breastfeeding within 1 hour. N=9,044 observations

Phase	Percentage point change in EBP adherence between intervention and control periods	Total EPB adherence during intervention period
Pilot 1	9.7% (-11%, 30%)	40% (23%, 56%)
Pilot 2	23% (17%, 28%)	37% (28%, 46%)
CRT[1]	33% (25%, 41%)	44% (39%, 50%)

[1]EBP adherence during CRT differs from what has been previously reported due to inclusion of 8, rather than 18, EBPs and because data is reported for entire post-intervention period rather than only for 2-month post-intervention and 12-month post-intervention periods.

Lesson 3: Create and capture variation in implementation intensity of components

If criteria for success are not satisfied, investigating relationships between individual implementation components and other variables in the TOC can help identity strategies for improving the intervention. These analyses require researchers to assess variation in implementation intensity, which can be viewed as the "strength" or the "dose" of each intervention component (58, 59). Variation in implementation intensity can arise from both planned and unplanned factors. Planned variation occurs when researchers assign study participants to receive different intensities of an intervention component, as is the case in multi-arm studies (60), factorial designs (7), or when researchers adapt the complex intervention across sequential phases of development. Planned variation can also occur when researchers phase in, phase out, or otherwise change the intensity of the components over time (58), as in a stepped wedge design (61). Unplanned variation in implementation intensity is often described in terms of fidelity, or the extent to which the content, frequency, duration, and coverage of implementation components delivered to participants deviates from the planned intervention (62-64). Although sources of variation in implementation intensity may be unplanned, they can be anticipated and measured. For example, researchers can document the dates of implementation component delivery, identify to whom components were delivered (and to whom they were not), and use self-reported or expert reviews to assess whether the delivery of the intervention occurred as planned (65).

Observing an association between the intensity an individual implementation component and relevant process outcomes identified in the TOC can provide evidence to support the effectiveness of that component (2). As with all observational research, the extent to which this

association reflects causal effects depends on the extent to which other confounders are accounted for (66-68). In the case of complex interventions, special care should be taken to adjust for the intensity of the remaining intervention components using either randomization (e.g. 69) or analytic approaches (e.g. 70, 71, 72). Researchers seeking to identify the effects of individual components should consider the extent to which various components' implementation intensities are correlated with each other. The more strongly two components are correlated, the more difficult it is to identify their independent effects. Strong correlations often arise from the study design. For example, problematic collinearity occurs when researchers simultaneously introduce, intensify, or diminish the intensity of multiple components in a single arm or phase of a study. Collinearity can also occur if a common factor, such as highly motivated leadership, simultaneously affects fidelity to multiple intervention components. To improve their ability to estimate the effectiveness of individual implementation components, researchers may wish to both create planned, uncorrelated variation in implementation intensity as part of the study's design and capture unplanned variation that arises in the field.

Lesson 3: Application to BetterBirth

Although the intensity of implementation package components varied across the BetterBirth development phases by design (Table 1.1), multiple components were simultaneously intensified in each phase. This practice created strong collinearities between individual components and prevented the identification of their individual effects. For example, the effect of having non-standardized leadership engagement could not be isolated from the effect of a three-day launch duration since each of these conditions appeared only in Pilot 1. Fidelity was

not systematically measured for any component. Despite these limitations, the BetterBirth intervention generated planned variation in the implementation intensity of coaching, which occurred frequently in the initial weeks of the intervention and gradually became less frequent over time. In addition to this planned source of variation, the BetterBirth team gathered data on the dates of the coaching visits, allowing us to assess unplanned variation that occurred when sites deviated from the intended coaching schedule. Unfortunately, due to the multi-collinearity of the remaining components, coaching is the only intervention package component whose effect can be individually analyzed.

In our TOC, we hypothesized that coaching would improve birth attendant motivation and ability, leading to increased EBPs adherence. We tested for the existence of this relationship by assessing the association between coaching intensity, defined for each infant as the number of coaching visits occurring at their site of birth in the 30 days prior to their birth, and overall EBP adherence. We observed a linear dose-response relationship between the number of coaching visits per month and overall EBP adherence (Figure 1.3). This association suggested that coaching was an effective intervention component. However, the model also illustrates that providing even 15 coaching visits per month would not have been sufficient to reach the criteria for success. This analysis suggested that other implementation components may have needed to be added or intensified for the intervention to be effective.

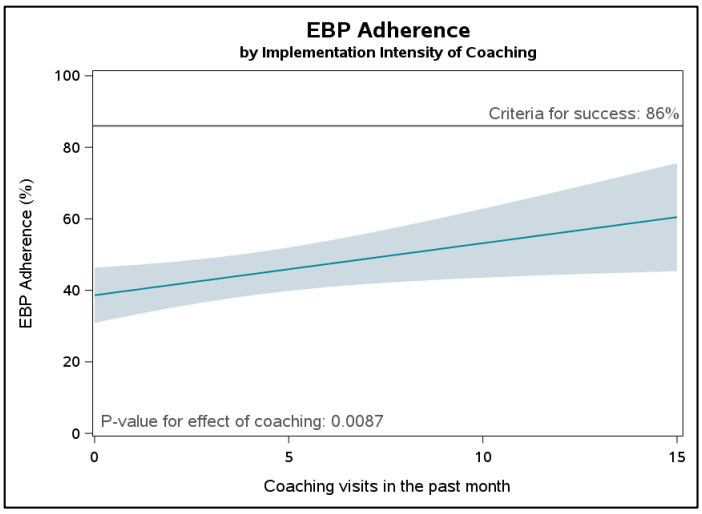

Figure 1.3 Dose-response relationship between EBP adherence and coaching intensity

DISCUSSION

Through our review of the development of the BetterBirth intervention and its associated trial, we identified three lessons learned that can help future researchers develop and evaluate complex interventions: 1) Develop a robust theory of change, 2) Define optimization outcomes and criteria for success, and 3) Create and capture variation in the implementation intensity of individual components. Our illustrative examples demonstrate how these lessons could have been applied to BetterBirth. Specifying a TOC prior to data collection could have promoted more complete data collection and generated additional opportunities to learn about the intervention. Identifying an optimization outcome that was a valid surrogate for maternal and neonatal health and comparing it against pre-defined criteria for success could have led to additional phases of intervention development prior to the full-scale trial. Finally, capturing and creating variation in implementation intensity for each implementation component could have helped identify which implementation components were effective and suggested potential areas for improvement of the implementation package.

These three lessons can be applied to both exploratory and more methodologically rigorous approaches to complex intervention development and evaluation. In fact, these lessons are highly compatible with the MOST framework, which also encourages researchers to begin with a well-developed theoretical framework and to only proceed to a full-scale trial after reaching some minimal effectiveness threshold (6, 9). They are also compatible with the LAGO design, which relies on variation in implementation intensity to estimate the effect of individual components and optimizes the intervention relative to some pre-specified criteria for success (10). Although we intend for these lessons to apply to the development and evaluations of complex interventions, many recommendations, including carefully selecting surrogate outcomes, defining success criteria, and adequately powering preliminary research, apply more generally to preliminary epidemiologic and public health studies, as has been previously noted (42, 43, 55).

We recommend that researchers operationalizing these lessons consult previous recommendations for developing theories of change (26, 27) and assessing fidelity (62, 64, 65). Surrogate outcomes have not been widely discussed in implementation science literature, so researchers should consult the epidemiologic literature for guidance (45). While researchers cannot know *a priori* whether an optimization outcome will be a valid surrogate for their primary outcome, using subject matter knowledge to critically evaluate the optimization outcomes' role in the context of the broader intervention can promote the selection of valid optimization outcomes (73). Because development phases for complex interventions are usually too small in size or short in duration to collect data on the primary outcome, we anticipate that process outcomes, rather the primary outcome itself, will typically be selected

as optimization outcomes. However, directly optimizing the primary outcome would avoid the risk of selecting an invalid surrogate and choosing inappropriate criteria for success. In these cases, criteria for success would be closely related to the minimum clinically important difference used for power calculations in a full-scale trail.

The selection of the BetterBirth intervention as a case study is both a strength and limitation of this paper. Because the BetterBirth Trial was a high-profile, large-scale study conducted by a research team with previous experience implementing behavioral change interventions (19, 20, 74, 75), it provides a realistic example of how complex interventions are currently being developed. The lessons learned from this case study are likely applicable to other teams currently developing and evaluating complex interventions. Additionally, the availability of quantitative data across three development phases allowed us to provide examples illustrating how quantitative analyses could be used to improve complex intervention development. However, because our lessons learned were identified using a single case study, they may not constitute an exhaustive list of factors researchers should consider when developing a complex intervention. Additionally, due to the limitations in the data, all quantitative results presented in the paper should be principally viewed as illustrative examples rather than valid estimates of causal effects. Although we identified areas in which the BetterBirth team's approach to intervention development could have been improved, the minimally structured multi-phase approach to complex intervention development approach used by the BetterBirth team is not unusual (see 5, 14, 76, 77 for similar examples). The popularity of this approach may stem from limits on the amount of time and resources researchers can dedicate to preliminary studies. Funders may wish to explore more flexible funding mechanisms with longer durations to ensure

that researchers have sufficient resources to fully optimize their evaluation prior to assessing its effectiveness.

When developing complex interventions, researchers must inevitably make difficult decisions that ultimately determine the success of the intervention. We feel that these decisions are more likely to be correct if researchers developing complex interventions first develop a theory of change and then test and refine this theory using quantitative analysis. We acknowledge that initial theories of change, criteria for success, and implementation intensity metrics may undergo substantial changes throughout the complex intervention development process. In addition to quantitative analyses, qualitative research is likely necessary for exploring unanticipated findings, contextualizing results, and generating new hypotheses. However, systematically documenting these processes and anticipating how they will impact subsequent quantitative analysis will promote learning throughout the intervention development process and give the final version of the complex intervention its best chance of succeeding in a full-scale trial.

CONCLUSION

As complex interventions become more common in health research, identifying strategies for developing and refining these interventions is critical. Theories of change, optimization outcomes, and implementation intensity metrics are generalizable strategies that can be used to improve the development and evaluation of complex interventions. By demonstrating the relevance of these strategies and how they can be applied in practice, we hope to encourage

the collection and use of data in a way that that promotes more effective development and evaluation of complex interventions.

Chapter Two: Coaching intensity, adherence to essential birth practices, and health outcomes in the BetterBirth Trial

ABSTRACT

Background: Globally, most maternal and perinatal deaths could be prevented through access to high-quality, facility-based care. Coaching is one implementation strategy that could improve the quality of care in primary-level birth facilities by promoting birth attendant adherence to essential birth practices (EBPs) known to reduce the risk of maternal and perinatal mortality. The appropriate intensity of coaching needed to promote and sustain behavior change is unknown. We used data from the BetterBirth Trial, which assessed the impact of a complex intervention designed to implement the World Health Organization's Safe Childbirth Checklist (SCC) in Uttar Pradesh, India, to investigate the relationship between coaching intensity, EBP adherence, and maternal and perinatal health outcomes.

Methods: For each birth during the study, we defined metrics reflecting multiple domains of coaching intensity, including coaching frequency (coaching visits delivered per month), cumulative coaching (total coaching visits delivered over the course of the intervention), and coaching fidelity (coaching delivered as scheduled). We considered coaching delivered at both the facility and the birth attendant levels. We assessed coaching intensity's association with birth attendant adherence to a set of 18 EBPs and with maternal and perinatal health outcomes using regression models.

Results: Coaching frequency was associated with increased EBP adherence. Delivering 6 coaching visits per month at the facility level was associated with adherence to 1.3 additional

EBPs (95% CI: 0.6, 1.9). High-frequency, high-coverage coaching at the birth attendant level was associated with greater improvements: providing 70% of birth attendants with at least one visit per month was associated with adherence to 2.0 additional EBPs (95% CI: 1.0-2.9). Neither cumulative coaching nor infidelity to the proscribed coaching schedule were associated with EBP adherence. Coaching was generally not associated with health outcomes, possibly due to the relatively small magnitude of association with EBP adherence.

Conclusions: Frequent coaching may promote birth attendant behavior change, especially if coaching is delivered with high coverage at the birth attendant level. However, the effects of coaching may not persist over the long term, suggesting that effective coaching-based interventions may require frequent coaching for longer periods.

Trial Registration: NCT02148952 registered on May 29th, 2014 at ClinicalTrials.gov

Keywords (3-10): India, Childbirth, Coaching, Checklist, Birth attendants, WHO Safe Childbirth Checklist, quality of care

INTRODUCTION

Rates of maternal and neonatal mortality in Low-and-Middle-Income Countries (LMICs) can be 10 times higher than in high-income countries (78, 79). Despite global increases in facility-based deliveries, progress in reducing these preventable deaths has been slower than expected due to poor quality of care in health facilities and poor adherence to evidence-based practices among birth attendants (80-84) . Improving the quality of care at birth facilities has the potential to avert 531,000 stillbirths, 1.3 million newborn deaths, and 112,000 maternal deaths each year (85). However, evidence-based strategies for improving the quality of care in birth facilities are lacking. Providing training alone can increase knowledge of evidence-based practices but does not necessarily translate into meaningful improvements in quality of care (86, 87). Consequently, additional implementation strategies are needed to improve the quality of intrapartum and postnatal care.

One promising implementation strategy to promote birth attendant behavior change is coaching. Coaching is a process that helps individuals use their existing skills, resources, and training to improve their performance and achieve personalized goals (88, 89). Unlike mentoring, which is focused more broadly on professional and personal development, coaching is task-oriented and performance-driven (90). To improve performance, coaches use multiple strategies, including modeling desired behaviors, providing supervision and feedback, and promoting problem solving (91). These strategies have been found to be effective at improving quality of care in LMICs across a variety of clinical areas (92, 93), including Integrated Management of Childhood Illness (94, 95), drug management and prescription practices (96,

97), primary care (98), malaria case management (99), voluntary male circumcision (100), and reproductive health (101, 102).

While some studies have reported associations between increased intensity of coaching-related activities and improved quality of care (97-99), the optimal coaching intensity needed to promote and sustain behavior change is unknown. Coaching intensity can be described across multiple domains, including frequency (e.g. two coaching visits per week); duration (e.g. six weeks of coaching); and cumulative dose, which reflects both the frequency and the duration of the intervention (e.g. two sessions per week for six weeks equals twelve cumulative visits) (59). Once the desired coaching regimen has been determined, coaching intensity can also be described in terms of fidelity, or the extent to which coaching is delivered as intended (63). Understanding which domains of coaching intensity are most strongly associated with quality of care improvements help identify coaching regimens that are optimized to promote behavior change and, ultimately, to improve health outcomes.

Domains of coaching intensity
adapted from Warren, Fey, and Yoder (2007)

Coaching form[1] – the coaching delivery method, including the coaches' identity and experience level (e.g. peer coaching, expert coaching) and the strategies used by the coach to generate behavior change (e.g. role playing, motivational support)

Coaching quality[1] – the ability of the coach to correctly and consistently use coaching strategies to generate behavior change

Coaching frequency - the number of coaching sessions delivered over a specific duration of time (e.g. two coaching visits per week)

Coaching duration[1] **–** the time period over which coaching is delivered (e.g. six weeks of coaching)

Cumulative coaching- the accrual of exposure to coaching over time, which is determined by both coaching frequency and coaching duration (e.g. two sessions per week for six weeks equals twelve cumulative visits)

[1]Describes a domain of coaching intensity that is not covered in this analysis

One coaching-based intervention designed to improve the quality of care provided to mothers and newborns during facility-based childbirth is the BetterBirth Program. Although this intervention did not reduce maternal morbidity or maternal and perinatal mortality in a recent matched-pair, cluster-randomized trial conducted in Uttar Pradesh, India, it did improve birth attendant adherence to many essential birth practices (EBPs) believed by experts to prevent or successfully manage complications during facility-based deliveries (56). In this paper, we use data from the BetterBirth Trial to assess coaching intensity's relationship with birth attendant adherence to EBPs and maternal and perinatal health outcomes. By investigating multiple dimensions of coaching intensity, we aim to provide insights into the optimal coaching regimen for future coaching-based interventions.

METHODS

Intervention

The BetterBirth Program was designed to promote the use of the World Health Organization's Safe Childbirth Checklist (SCC), a 28-item tool intended to assist birth attendants in performing EBPs. Coaching has been recommended as a core component of SCC implementation packages since the checklist's initial development (18) and was a major feature of the BetterBirth Program's multicomponent implementation package. The BetterBirth Program used an Engage-Launch-Support model, which has been previously described in detail (16, 17, 103). Briefly, district and facility level leadership were introduced to the BetterBirth Program and engaged in identifying and addressing local needs. Then, an educational and motivational launch event was held at each facility to introduce birth attendants to the SCC and train them on its use. Finally,

ongoing coaching and data feedback were used to support behavior change. Peer-to-peer coaching occurred at both the birth attendant and facility leadership levels with birth attendants, who were primarily nurses, being coached by trained nurses and facility leadership being coached by physicians or health professionals. Coaching followed an "Opportunity-Ability-Motivations-Supplies" framework adapted from previous behavior change models (104, 105). In this framework, coaches motivated birth attendant behavior change; collected data and provided feedback on current adherence to EBPs; identified existing opportunity-, ability-, motivation-, or supply-related barriers to EBP adherence; and engaged in group problem solving to address these barriers. Coaches did not provide additional technical training but could use strategies like role-playing or advocating for additional training opportunities to address ability-related barriers. Birth attendant coaching was scheduled to occur twice per week during the first through fourth months of the intervention, once per week during the fifth and sixth months of the intervention, once every two weeks during the seventh month, and once per month in the eighth month for a total of 43 visits. Facility leadership coaching followed a similar, but less intensive schedule for a total of 23 visits. The same coaching schedule was proscribed to all facilities, regardless of the facility's delivery load or birth attendant staff size. To promote long-term sustainability, each site designated a Childbirth Quality Coordinator who was intended to serve as a long-term, facility-based coach.

Trial design and study setting

This implementation package was evaluated in a matched-pair, cluster-randomized trial that enrolled 120 primary-level health facilities in Uttar Pradesh, India, a region with high maternal (258/100,000 live births) and neonatal (49/1,000 live births) mortality (106). All facilities

enrolled in the BetterBirth Trial were required to conduct at least 1,000 deliveries per year, have at least 3 birth attendants trained at the level of auxiliary nurse midwife or higher, have no concurrent quality improvement or research programs, and have district and facility leadership who were willing to participate. Eligible facilities were matched on baseline characteristics and randomized within pairs to receive either the coaching-based intervention or the current standard of care. Roll out of the intervention was staggered across six geographically-defined research hubs centered in the urban areas of Agra, Gorakhpur, Lucknow, Meerut, and Varanasi. Full details on study design and data collection, including sample size calculations, can be found elsewhere (22).

Data collection

At each facility, registers were used to document the date of admission for each birth as well as any instances of mortality and morbidity occurring at the facility. Data on seven-day health outcomes were obtained using a call center, which contacted mothers and their families via mobile phone between 8 and 42 days postpartum, followed by home-visits if neither the woman nor a family member could be reached by phone after 22 days postpartum (107). In a convenience sample of births occurring in 30 facilities (15 intervention, 15 control) additional data were collected on birth attendant EBP adherence, which was directly observed by trained independent nurses and recorded using standardized data collection tools. These 30 facilities were located in the Lucknow hub, which was located in central Uttar Pradesh. These independent observers did not intervene in clinical care. EBP observations occurred during three distinct periods of delivery: on admission to facility, just before pushing, and within one hour after birth. However, practical considerations related to the timing and duration of labor

prevented all births from being continuously observed from admission through discharge such that not all EBPs were observed for each birth. For intervention facilities, the date of each coaching visit as well as the identity of the birth attendants who were coached during that visit were recorded by the nurse coaches.

Outcomes

We considered two types of outcomes: birth attendant EBP adherence and maternal and perinatal health outcomes. EBP adherence was defined as the number birth practices that were successfully completely by a birth attendant out of 18 practices that should be completed for all mother-infant dyads (Table 2.1) (56). Births were eligible for inclusion in our EBP analysis if they occurred at one of the 15 intervention facilities where EBP adherence data was collected, if they occurred after the start of coaching at that facility, and if the birth been directly observed during admission to facility, just before pushing, and within one hour after birth such that adherence to all 18 practices was recorded. Following the main trial, our primary health outcome was a composite outcome of events occurring within 7 days after delivery which included severe maternal morbidity, defined as self-reported complications including seizures, loss of consciousness for more than one hour, fever with foul-smelling vaginal discharge, hemorrhage, or stroke; maternal mortality; or perinatal mortality, defined as stillbirth or death within the first seven days of life. A secondary composite health outcome consisting of only seven-day maternal or perinatal mortality was also considered (56). Births were included in the health outcomes analysis if they if they occurred in an intervention facility after the start of coaching, if mothers had consented to follow-up, and if data on seven-day outcomes had been obtained. Because the timing of direct observations of birth attendant

adherence to EBPs (which occurred 0-8 and 13-17 months after the start of coaching) differed somewhat from the timing of call center activities (which continued from 0-13 months after the start of coaching), the EBP adherence sample is not a subset of health outcomes sample. However, some births appear in both samples.

Table 2.1 Eighteen Essential Birth Practices (EBPs). Independent observers assessed birth attendant adherence to EBPs, but not the technical skill or quality with which the EBP was performed.

At Admission	**Before Pushing**	**After Birth**	**Any time**
Partograph started	Hand hygiene	Oxytocin administered within 1 minute	Maternal temperature taken
Birth companion present	Clean towel available	Birth companion present	Maternal blood pressure taken
	Clean blade available	Baby weighed	
	Cord tie available	Baby temperature taken	
	Mucus extractor available	Skin-to-skin warming initiated	
	Neonatal bag available	Skin-to-skin warming maintained for 1 hour	
	Clean pads available	Breast feeding initiated	

Coaching Intensity

For each birth, we calculated metrics that reflected multiple domains of coaching intensity, including coaching frequency, cumulative coaching, and coaching fidelity. These metrics were based on the dates of the peer-to-peer birth attendant coaching visits that had occurred at a given facility prior to each birth. For coaching frequency, we assigned each birth a coaching intensity equal to the number of coaching visits occurring at that facility in the 30 days prior to the date of admission (visits in the past month). Because we hypothesized that the impact of coaching on birth-related outcomes would be stronger when we considered the intensity of coaching provided to the birth attendants conducting the delivery rather than to the facility as a whole, we also created coaching frequency metrics that reflected coaching delivered at the birth attendant level. In our data, it was not possible to identify which birth attendants conducted specific deliveries, so we created birth attendant (BA) level coaching metrics that were aggregated at the facility level, including average number of visits in the past 30 days among birth attendants (mean visits per month per BA), the proportion of birth attendants who had received at least one coaching visit in the past month (% BAs receiving ≥1 visit in past month), and the standard deviation of coaching visits in the past month among birth attendants (standard deviation in visits among BAs in past month). We hypothesized that facilities would experience greater benefits from coaching if birth attendants had, on average, a greater number of visits in the past month (mean visits per month per BA), higher coaching coverage (% BAs receiving ≥1 visit in past month), and a more equal distribution of coaching visits among birth attendants (lower standard deviation in visits among BAs in the past month). All birth-attendant level coaching metrics were calculated under the assumption that the birth

attendants listed in the coaching database reflected a complete list of birth attendants employed by the facility over the course of the intervention. These metrics also did not consider staff turnover, which was assumed to be minimal over the intervention period. We also explored coaching frequency metrics calculated over a one-week, rather than a one-month, time horizon. However, since these two sets exposures produced similar results, we have presented only the results for the one-month time horizon. Results for the one-week time horizon can be found in the Supplemental Materials.

For cumulative coaching, we assigned each birth a coaching intensity equal to the total number of coaching visits accrued at the facility between the start of program and their date of admission (total visits). As with coaching frequency, we believed that a) coaching delivered at the birth attendant level would be have a greater impact than coaching delivered at the facility level and b) facilities with more equal coverage of coaching among birth attendants would experience greater benefits from coaching. Therefore, for each birth we also calculated the mean number of visits accrued among birth attendants between the start of program and the admission date (mean visits per BA), the proportion of birth attendants who had received at least 10 coaching visits (% BAs receiving ≥10 visits), and the standard deviation of coaching visits among birth attendants (standard deviations in visits among BAs).

Coaching fidelity was defined according to the proscribed coaching schedule of attaining at least two visits per week during the first four months of the intervention, at least one visit per week during the fifth and sixth months, at least one visit every two weeks during the seventh month, and at least one visit per month during the eighth month. Current coaching infidelity was a binary variable reflecting whether the date of admission occurred on a day when the

facility had deviated from this schedule. Cumulative coaching infidelity reflected the total number of non-adherent days accrued between the start of program and the date of admission. For example, if a facility had been three days late for its first coaching visit and four days late for its second coaching visit, then subsequent births would receive a cumulative coaching infidelity value of seven.

Statistical methods

Because the BetterBirth Program proscribed high-frequency coaching early in the intervention and gradually reduced the frequency of coaching over time, there were strong correlations among coaching frequency metrics, cumulative coaching metrics, and time since the start of the intervention. We explored these correlations graphically and using Spearman correlation coefficients. To assess associations between each metric of coaching intensity and our outcomes of interest, we used generalized linear models and accounted for clustering at the facility level by estimating standard errors using the empirical variance with an exchangeable working covariance structure (23). For EBP adherence, we estimated the change in the number of EBPs adhered to associated with each coaching intensity metric using an identity link and a normal distribution. For binary health outcomes, we estimated the risk ratios associated with each coaching intensity metric using a log link and a binomial distribution (108). Because coaching metrics had very different ranges (e.g. total coaching ranged from 1 to 47 while % of BA receiving ≥10 visits ranged from 0 to 1), we report effect sizes associated with increasing each continuous coaching metric from its 25th percentile to its 75th percentile, or by one interquartile range (IQR). These percentiles were calculated in the health outcomes dataset. Where relevant, we also report effect sizes for a one-unit increase. For all models, we used

score tests to assess the statistical significance of model parameters (109). Our primary models adjusted for facility-level covariates, including research hub location; being located in a high-priority district, which is a designation used by the Indian government to identify districts with a high overall burden of mortality; distance to district hospital in kilometers; and number of skilled birth attendants at that facility. At the birth level, models also adjusted for whether or not the birth occurred on the same day as a coaching visit. We fit models for each coaching metric separately and also used stepwise regression to assess whether multiple coaching metrics should be included in the same model based on an $\alpha \leq 0.05$ criteria for model entry and exit. Because the effects of behavior change interventions often fade over time (110), a phenomenon which would bias results against cumulative coaching metrics and in favor of coaching frequency metrics, in a secondary set of models we additionally adjusted for months since the start of the intervention. We tested for potential non-linear relationships between months since the start of the intervention and our outcomes of interest using restricted cubic splines (24) selected using a publicly available SAS macro (111). Finally, we assessed whether the association between coaching intensity and our outcomes of interest changed over the course of the intervention by adding an interaction between each coaching metric and months since the start of the intervention to our models. Because of the strong collinearities between coaching metrics and months since the start of the intervention, several models produced statistically-significant interaction terms that were not interpretable. Therefore, to ensure interpretability, we reported results for these interaction models only if both the time-by-coaching interaction term and the overall effect of coaching based on the joint null hypothesis

that both the main effect of coaching and its interaction term were zero, were statistically significant at the $\alpha=0.05$ level.

RESULTS

Study population

Data on EBP adherence at intervention facilities was collected for 3,283 births. We excluded 262 births that occurred before the start of the coaching intervention and 938 that were not observed for all three pause points for a final sample of 2,083 births. Health outcomes data at intervention facilities was collected among 83,166 births. We excluded six deliveries referred in from another facility, 436 deliveries that occurred outside of the facility, five women admitted for abortion, 352 births that occurred before the start of coaching, 1,868 births where patients did not consent to follow-up, and 265 births that were lost to follow-up for a final sample of 80,234 births (Figure 2.1). An additional 457 births lacked complete data on maternal morbidities and were excluded from analyses of the primary composite outcome. The EBP adherence and health outcome samples overlapped by 1,100 births and shared many similarities (Table 2.2); however, facilities in the EBP adherence sample came exclusively from the Lucknow hub, located in the center of the state, and were more likely to be in a high priority district. Due to differences in the timing of data collection in the two samples, births in the EBP adherence sample were less likely to have occurred on a coaching day.

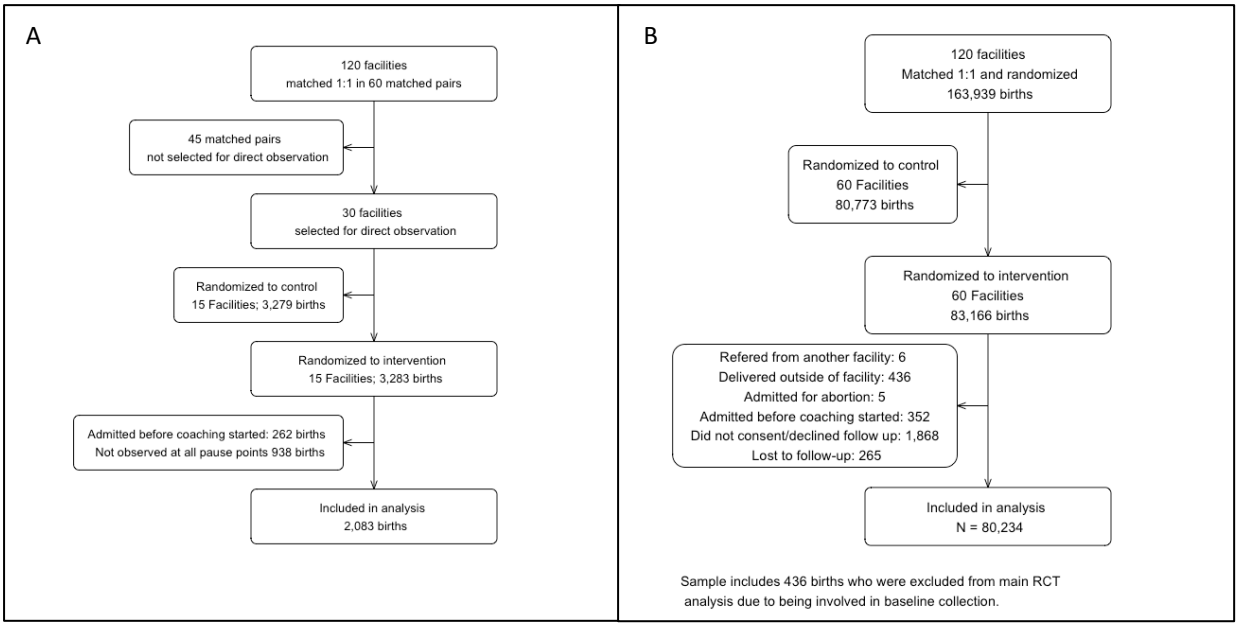

Figure 2.1 Study populations for EBP adherence (A) and health outcomes (B) analyses

Table 2.2 Descriptive statistics for EBP adherence and health outcomes study populations

	EBP Adherence Sample	**Health Outcomes Sample**
	N (%) / Mean (SD)	**% / Mean (SD)**
Facility-level variables	**N=15**	**N=60**
Research hub		
Agra	---	9 (15%)
Gorakhpur	---	11 (18%)
Lucknow	15 (100%)	19 (32%)
Meerut	---	7 (12%)
Varanasi	---	14 (23%)
High priority district	7 (47%)	7 (12%)
Distance to district hospital (km)	29 (12)	30 (14)
Number of skilled birth attendants	4.5 (1.1)	4.4 (1.2)
Annual delivery load	1,795 (468)	1,599 (435)
Birth Level variables	**N=2,083**	**N=80,234**
Birth occurred on coaching day	107 (5.1%)	7,533 (9.4%)
Months since intervention started at facility	8.5 (5.8)	6.7 (2.8)
EBP Adherence (out of 18 practices)	12 (2.4)	---
Primary Composite[1]	---	12,062 (15%)
Secondary Composite	---	3,907 (4.9%)

[1]457 births are missing data on maternal morbidity and are therefore missing data on the primary composite outcome

Coaching intensity

Fidelity to the coaching schedule was very high. By the end of the intervention, 53/60 (88%) facilities reached the target of 43 total coaching visits, six (10%) reached 42 visits, and one facility reached 37 visits. However, fidelity at the facility level did not necessarily translate into high coverage at the birth attendant level. While birth attendants attained, on average, 10 coaching visits by the end of the intervention, 34% attained fewer than 5 visits. As would be expected based on the proscribed coaching schedule, cumulative coaching metrics increased with months since the start of the intervention while coaching frequency metrics decreased over time (Figure 2.2). Cumulative coaching measures were positively associated with time since intervention (ρ: 0.36 to 0.87) and with each other (ρ =0.30 to 0.86) while coaching frequency metrics were negatively associated with time since intervention (ρ=-0.82 to -0.97) and positively associated with each other (ρ: 0.79 to 0.98) (Supplemental Materials Table 2.5).

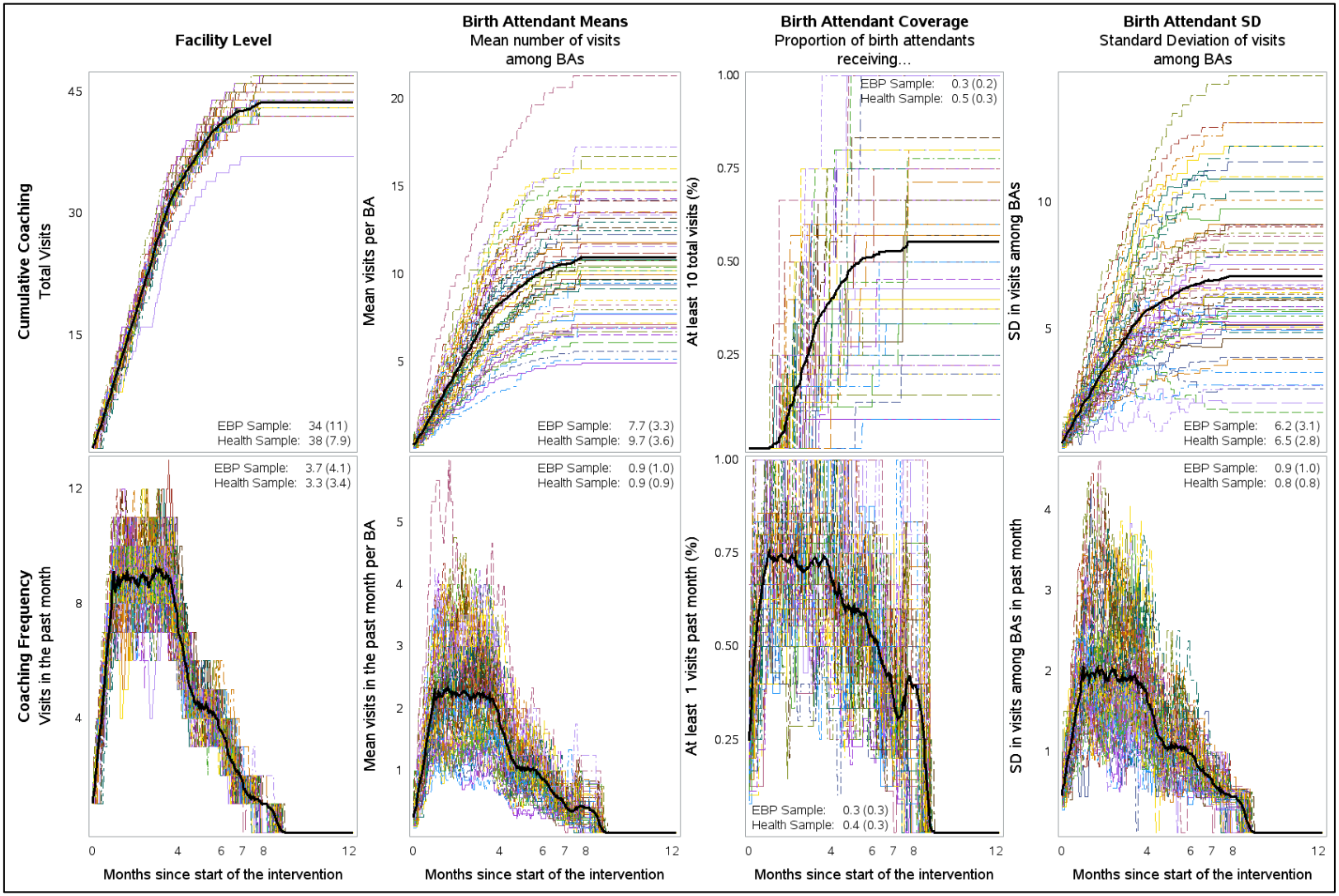

Figure 2.2 Coaching intensity metrics over time. Each panel provides the mean and (standard deviation) for each coaching metric in the EBP adherence and health outcome samples.

EBP Adherence

In our primary model, all four coaching frequency metrics were strongly associated with increased EBP adherence (Table 2.3). On average, providing a facility with six coaching visits per month was associated with adhering to an additional 1.3 EBPs (95% CI 0.6, 1.9). The association between EBP adherence and coaching frequency was even stronger if coaching coverage was high among BAs: providing 70% of BAs with at least one visit per month was associated with adherence to 2.0 additional EBPs (95% CI: 1.0, 2.9) and achieving 100% coverage of monthly visits was associated with adherence to 2.8 additional EBPs (95% CI: 1.4-4.2). However, no cumulative coaching or coaching infidelity metrics were significantly associated with EBP adherence. Our stepwise selection procedure did not identify a model that included multiple coaching metrics.

When we included months since the start of the intervention in our model, we did not detect any non-linear effects of time. After adjusting for time since the start of the intervention mean visits in the past month per BA and % of BAs receiving ≥1 visit in the past month, two coaching frequency metrics that were both assessed at the birth attendant level, remained significantly and positively associated with increased EBP adherence. While non-significant, adjusting for time since start of the intervention also resulted in the cumulative coaching metrics becoming non-significantly associated with increased EBP adherence.

Table 2.3 Association between coaching intensity and EBP adherence. Effects are reported for both a one-unit increase and for increasing each continuous coaching metric from its 25th percentile to its 75th percentile, or by one interquartile range (IQR). Results are from a generalized linear with an identity link. Standard errors are estimated using the empirical variance with an exchangeable working covariance structure to account for clustering at the facility level. (N=2,083 births)

		Model 1[1]			Model 2[2]		
Coaching Domain	**Units in IQR increase**	**Δ in practices adhered to associated with one-unit increase (95% CI)**	**Δ in practices adhered to associated with IQR increase (95% CI)**	**p-value**	**Δ in practices adhered to associated with one-unit increase (95% CI)**	**Δ in practices adhered to associated with IQR increase (95% CI)**	**p-value**
Coaching Frequency							
Visits in the past month	6.0	0.2 (0.1, 0.3)	1.3 (0.6, 1.9)	<.01	0.2 (0.0, 0.4)	1.0 (-0.1, 2.2)	0.10
Mean visits in the past month per BA	1.3	1.0 (0.6, 1.4)	1.2 (0.7, 1.8)	<.01	0.9 (0.2, 1.6)	1.2 (0.3, 2.1)	0.01
% of BAs receiving ≥1 visit in past month	70%	2.8 (1.4, 4.2)	2.0 (1.0, 2.9)	0.01	3.4 (1.0, 5.8)	2.4 (0.7, 4.0)	0.03
Standard deviation in visits among BAs past month	1.3	0.9 (0.5, 1.4)	1.2 (0.6, 1.8)	0.01	0.7 (0.0, 1.5)	1.0 (0.0, 1.9)	0.08
Cumulative coaching							
Total visits	8.0	-0.0 (-0.1, 0.0)	-0.4 (-0.8, 0.1)	0.09	0.1 (0.0, 0.1)	0.6 (0.3, 0.9)	0.07
Mean visits per BA	5.3	-0.2 (-0.4, 0.0)	-1.0 (-2.1, 0.1)	0.09	0.2 (0.0, 0.4)	1.0 (0.0, 2.0)	0.21
% of BAs receiving ≥10 visits	40%	-3.0 (-5.9, -0.1)	-1.2 (-2.4, 0.0)	0.12	0.3 (-3.5, 4.1)	0.1 (-1.4, 1.6)	0.89
Standard deviation in visits among BAs	3.5	-0.2 (-0.4, 0.1)	-0.6 (-1.5, 0.2)	0.12	0.3 (0.0, 0.5)	0.9 (0.2, 1.7)	0.08
Coaching infidelity							
Current coaching infidelity	NA[3]	0.3 (-0.7, 1.3)	---[3]	0.55	-0.5 (-1.3, 0.4)	---[3]	0.27
Cumulative coaching infidelity	12	-0.0 (-0.1, 0.0)	-0.5 (-1.0, 0.1)	0.08	0.1 (0.0, 0.1)	0.8 (0.0, 1.6)	0.11

[1]Adjusted for whether the facility was in a high-priority district, distance to district hospital, facility staff size, facility delivery load, whether birth occurred on the same day as a coaching visit. [2] Adjusted for everything in Model 1 plus months since start of the intervention. [3]Because current coaching infidelity is a binary outcome, we report the effect for infidelity vs. no infidelity, rather than for a one IQR increase

When we included an interaction term between coaching intensity metrics and time since the start of the intervention, the effect of coaching was found to vary over time for only one coaching metric, mean coaching visits per BA (Supplemental Materials Table 2.6). This cumulative coaching measure was associated with increased EBP adherence in the early months of the intervention when coaching occurred very frequently, but the positive association did not persist after coaching visits ceased (Figure 2.3)

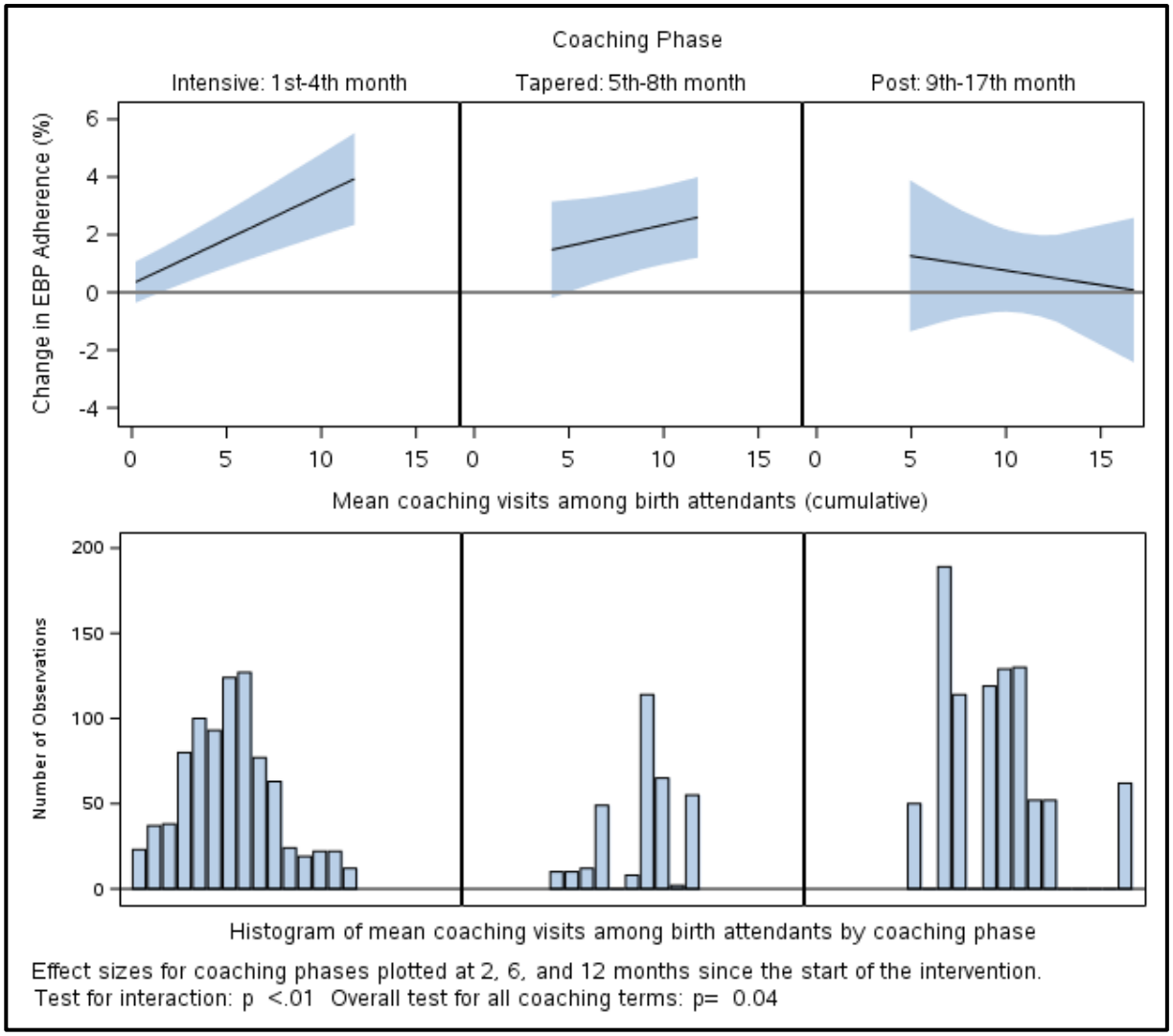

Figure 2.3 Effect modification of the association between mean coaching visits among birth attendants (cumulative) and EBP adherence over months of the intervention (N=2,083)

Health outcomes

In general, coaching was not associated with health outcomes (Table 2.4). In our primary model, coaching infidelity was associated with increased risk of the primary composite outcome, which reflected maternal morbidity, maternal mortality, and perinatal mortality; but this result attenuated after adjusting for months since the start of the intervention. After adjusting for months since the start of the intervention, we also observed a significant association between average visits per BA and increased risk of the primary composite outcome (RR: 1.10, 95% CI: 1.03, 1.18). However, because this model also estimated an implausibly strong 18% reduction in the risk of mortality or morbidity over the course of a year (RR: 0.82, 95% CI: 0.71, 0.95), this association and likely reflects strong correlations between time and coaching rather than a true adverse effect of coaching. Our stepwise selection procedure did not identify a model that included multiple coaching metrics, and no significant interactions were detected

Table 2.4 Risk ratios for the association between coaching and health outcomes. Effects are reported for increasing each continuous coaching metric from its 25th percentile to its 75th percentile, or by one interquartile range (IQR). Results are from a generalized linear model with a log link and binomial distribution. Standard errors are estimated using the empirical variance with an exchangeable working covariance structure.

		Primary Composite Maternal morbidity or maternal or infant mortality (n/N=12,062/79,777)				**Secondary Composite** Maternal or infant mortality (n/N=3,907/80,234)			
		Model 1[1]		**Model 2[2]**		**Model 1[1]**		**Model 2[2]**	
Coaching domain	**Units in increase**	**RR (95% CI)**	**p-value**	**RR (95% CI)**	**p-value**	**RR (95% CI)**	**p-value**	**RR (95% CI)**	**p-value**
Coaching Frequency									
Visits in the past month	6.0	1.03 (0.99, 1.07)	0.14	1.03 (0.94, 1.13)	0.49	0.98 (0.92, 1.05)	0.61	1.01 (0.88, 1.15)	0.91
Mean visits in the past month per BA	1.3	1.03 (1.00, 1.06)	0.10	1.02 (0.97, 1.08)	0.36	0.98 (0.93, 1.04)	0.54	0.99 (0.92, 1.07)	0.84
% of BAs receiving ≥1 visit in past month	70%	1.05 (1.00, 1.10)	0.05	1.06 (0.98, 1.15)	0.15	1.00 (0.92, 1.09)	0.99	1.07 (0.93, 1.22)	0.36
Standard deviation in visits among BAs in past month	1.3	1.02 (0.98, 1.05)	0.30	0.99 (0.94, 1.05)	0.82	1.00 (0.95, 1.06)	0.98	1.06 (0.99, 1.14)	0.11
Cumulative coaching									
Total visits	8.0	0.99 (0.97, 1.02)	0.50	1.02 (0.98, 1.05)	0.42	1.01 (0.97, 1.05)	0.59	1.00 (0.94, 1.06)	0.93
Mean visits per BA	5.3	1.01 (0.96, 1.07)	0.70	1.10 (1.03, 1.18)	0.03	1.07 (0.98, 1.18)	0.18	1.10 (0.98, 1.25)	0.16
% of BAs receiving ≥10 visits	40%	1.00 (0.96, 1.05)	0.87	1.04 (0.99, 1.09)	0.14	1.04 (0.96, 1.13)	0.32	1.04 (0.94, 1.15)	0.42
Standard deviation in visits among BAs	3.5	1.00 (0.95, 1.06)	0.89	1.04 (0.98, 1.11)	0.25	1.04 (0.96, 1.13)	0.34	1.04 (0.96, 1.14)	0.37
Coaching infidelity									
Current coaching infidelity[3]	1	1.06 (1.01, 1.12)	0.04	1.05 (1.00, 1.11)	0.08	0.97 (0.86, 1.09)	0.58	0.98 (0.87, 1.10)	0.69
Cumulative coaching infidelity	12	1.00 (0.96, 1.05)	0.85	1.06 (0.99, 1.13)	0.10	1.03 (0.98, 1.09)	0.23	1.05 (0.99, 1.12)	0.12

[1]Adjusted for hub name, whether the facility was in a high-priority district, distance to district hospital, facility staff size, facility delivery load, whether birth occurred on the same day as a coaching visit

[2] Adjusted for everything in Model 1 plus months since start of the intervention

[3] Because current coaching infidelity is a binary outcome, we report the effect for infidelity vs. no infidelity, rather than for a one IQR increase

DISCUSSION

Our analysis suggests that in the BetterBirth Trial, coaching frequency was associated with increased EBP adherence. Associations between coaching frequency and EBP adherence tended to be stronger when considering coaching delivered at the birth attendant, rather than the facility, level. In contrast, cumulative coaching was generally not associated with EBP adherence. However, when we adjusted for time since the start of the intervention, cumulative coaching metrics became non-significantly positively associated with and EBP adherence, and, when we allowed the effect of coaching to change over time, one cumulative coaching metric, mean visits per BA, was significantly associated with increased EBP adherence in the early months of the intervention but not after coaching visits ceased. Because the BetterBirth coaching schedule created strong correlations among coaching metrics, it is difficult to isolate the independent effects of coaching frequency and cumulative coaching. Despite this limitation, our analyses suggest that high-frequency, high-coverage coaching can prompt birth attendant behavior change. However, the positive effects of coaching diminish over time and sustaining the effects of a coaching-based intervention may require high-frequency, high-coverage coaching over a long-term period.

Health outcomes were generally not associated with coaching. This lack of association may reflect the fact that the magnitude of the association between coaching and total EBP completion was relatively modest. This small absolute change may not have produced sufficient improvements in the quality of care to impact health outcomes. We did observe an increased probability of experiencing maternal morbidity, maternal mortality, or perinatal mortality on days when sites had deviated from the intervention's proscribed coaching schedule. Because

we did not observe evidence of protective effects of coaching in any other models, this association likely does not reflect direct benefits of coaching. Instead, it may suggest that sites that were unable to adhere to the coaching schedule were also experiencing other structural issues, such as poor leadership or inaccessibility, that placed mothers and infants at risk of harm.

While previous papers have reported dose-responses between coaching intensity and coaching effectiveness (97-99), to our knowledge, this is the first paper to simultaneously investigates multiple domains of coaching intensity. Consequently, there was relatively little guidance on how to define coaching intensity metrics. Contrary to initial expectations, greater standard deviations in visits among BAs in the past month, which we expected to reflect the unequal distribution coaching among birth attendants and be associated with worse outcomes, was associated with improved EBP adherence. This metric was strongly positively associated with the other coaching frequency metrics (ρ=0.79-0.98) and therefore likely did not perform well as an independent metric of unequal coaching among birth attendants. Similarly, cumulative coaching infidelity and cumulative standard deviation in visits among birth attendants, both of which were hypothesized to have adverse effects, were highly correlated with and produced results similar to those cumulative coaching metrics believed to have beneficial effects. In many coaching-based interventions, cumulative infidelity to the coaching schedule and disparities in the distribution of coaching among health care workers would be expected to increase over time. Consequently, cumulative coaching infidelity and standard deviation-based coaching metrics may exhibit problematic correlations with other coaching metrics in many settings and may therefore be generally difficult to interpret as metrics of coaching intensity.

Our finding that frequent coaching is associated with moderate improvements in EBP adherence is similar to previous reports. A recent meta-analysis found that strategies used by coaches, including supervision, training, and group problem solving, are associated with improving healthcare worker performance by 1, 6.4, and 13.6 percentage points, respectively (93). Our models suggest that providing a facility with 6 coaching visits per month is associated adherence to 1.3 additional EBPs, which corresponds to a 7.2 percentage-point increase on our 18-point EBP adherence scale. While the magnitude of the association between coaching and EBP adherence is relatively small, providing high-frequency, high-coverage coaching may be able to make a more dramatic difference Our results also parallel findings related to Low Dose, High Frequency (LDHF) training, which combines brief onsite trainings with short, frequent practice sessions and has resulted in positive behavior change among birth attendants and improved maternal and child health outcomes (112, 113). While LDHF models emphasize the acquisition of new skills through training and practice rather than coaching, both the LDHF model and our results suggest that frequent contact with health care workers may be necessary to improve quality of care.

In this study, all facilities were front-line health facilities receiving peer-to-peer coaching provided by external coaches according to the "Opportunity-Ability-Motivations-Supplies" framework. Consequently, we do not know the extent to which the relationships observed in this analysis can be generalized to other forms of coaching. Other programs that have used coaching as a Safe Childbirth Checklist implementation strategy have reported extremely variable EBP adherence at endline (Median: 64%, Range: 32-93%) (21, 114-118). In general, these interventions have not specified behavior change models nor have they provided full

details on the frequency or duration of coaching delivered. Consequently, it is difficult to determine the extent to which observed differences in the effectiveness of these interventions result from contextual differences between study settings or differences in their coaching intensity. Furthermore, some of these studies relied primarily on internal coaches recruited from within intervention facilities (115, 118). While the intensity of external coaching interventions can be evaluated using dates of coaching visits, this approach would not apply to internal coaches who are embedded within the intervention facilities and may therefore engage in coaching for variable amounts of time each day. Alternative approaches of assessing coaching intensity, such as time-motion studies, may be more appropriate for evaluating the intensity of internal coaching strategies (119).

Our analysis has several limitations. As discussed above, the BetterBirth Program's coaching schedule created strong correlations among coaching intensity metrics, which complicates the interpretation of some findings. Second, although we sought to minimize bias by adjusting for facility-level characteristics, residual confounding is possible if unmeasured facility or birth attendant characteristics impacted coaches' behavior. For example, reports suggest that in order to minimize travel time coaches would provide difficult-to-reach facilities with visits on back-to-back days. If less accessible facilities experienced worse outcomes, we would expect this practice to bias our results against coaching frequency metrics. Coaches may have also been more likely to provide coaching to the facilities or birth attendants who were most motivated and receptive to their help, which we would expect to bias our results in favor of coaching. Finally, because we were unable to link individual birth attendants to individual births in our dataset, we assessed coaching delivered at the birth-attendant level using aggregated

metrics that did not consider staff turnover within the health facility. We would expect this measurement error to underestimate the direct benefits of providing coaching to individual birth attendants.

Despite these limitations, our research provides several practical insights for those seeking to implement or study future coaching-based interventions. First, high fidelity to a facility-centric coaching schedule will not necessarily translate to good coaching coverage among health care workers. Interventionists should specify and monitor coaching delivered at both the facility and the health-care-worker levels. Second, our study suggests that high-frequency coaching can impact health worker behavior. Future interventionists may wish to explore cost-effective methods for maintaining high-frequency coaching over longer periods of time, such as recruiting internal coaches or combining in-person coaching visits with remote coaching conference calls. Third, identifying optimal coaching regimens requires designing interventions that have uncorrelated variation in coaching frequency and cumulative coaching. This could be achieved, for example, by conducting a three-armed trial with one control arm, one arm receiving evenly spaced coaching visits over a set duration of time, and a third arm receiving the same total number of coaching visits delivered over the same duration of time but following a tapered schedule similar to the BetterBirth Program's. Finally, statistically significant improvements in quality of care indicators do not necessarily translate into meaningful improvements in health outcomes. In general, changes in quality of care indicators will be more likely to predict improvements in health outcomes if researchers choose a quality of care indicators have been empirically demonstrated to have a causal relationship with the health outcome and if the magnitude of the change is clinically meaningful on the absolute scale (e.g.

EBP adherence increased by 20 percentage points from 60% to 80%) rather than on the relative scale (e.g. EBP adherence doubled from 5% to 10%) (120). We recommend that future researchers consider whether the quality of care improvements observed in their study are large enough to plausibly generate meaningful improvements in the health outcomes. If not, more frequent coaching, additional implementation strategies, or both may be required to produce the desired health impact.

CONCLUSIONS

Frequent coaching was associated with increased adherence to essential birth practices among birth attendants in the BetterBirth Trial. This association tended to be stronger for coaching delivered at the birth attendant level compared to coaching delivered at the facility level. Cumulative coaching metrics were not associated with essential birth practice adherence, suggesting that the effects of coaching may not persist over time. Effective coaching-based interventions may require more frequent, high-coverage coaching for longer periods. Coaching was generally not associated with health outcomes, suggesting that additional coaching and other implementation strategies may be needed to achieve the desired health impact.

SUPPLEMENTAL MATERIALS

Table 2.5 Spearman correlation coefficients (ρ) among coaching metrics and intervention duration. Top numbers refer to EBP adherence dataset (N=2,083) and bottom to health outcomes dataset (N=80,234). Groups of cumulative coaching and coaching frequency variables are bolded.

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**O**
Months since the start of the intervention (A)	1.00	-.85	-.87	-.85	-.87	-.82	-.81	-.81	-.81	0.81	0.65	0.57	0.60	-.23	0.77
	1.00	-.97	-.93	-.82	-.92	-.83	-.79	-.77	-.80	0.87	0.50	0.42	0.36	-.23	0.71
Visits in the past month (B)		1.00	**0.96**	**0.92**	**0.96**	**0.90**	**0.88**	**0.86**	**0.88**	-.78	-.55	-.50	-.43	0.16	-.71
		1.00	**0.96**	**0.85**	**0.95**	**0.86**	**0.82**	**0.80**	**0.83**	-.85	-.47	-.38	-.33	0.19	-.71
Mean visits in the past month per BA (C)			1.00	**0.95**	**0.98**	**0.89**	**0.91**	**0.89**	**0.90**	-.80	-.48	-.45	-.37	0.17	-.70
			1.00	**0.93**	**0.94**	**0.82**	**0.83**	**0.81**	**0.82**	-.80	-.32	-.25	-.28	0.19	-.67
% of BAs receiving ≥1 visit in past month (D)				1.00	**0.89**	**0.86**	**0.86**	**0.87**	**0.84**	-.78	-.50	-.50	-.49	0.18	-.68
				1.00	**0.79**	**0.71**	**0.73**	**0.73**	**0.71**	-.69	-.23	-.16	-.30	0.17	-.57
Standard deviation in visits among BAs in past month (E)					1.00	**0.87**	**0.88**	**0.85**	**0.89**	-.79	-.51	-.46	-.35	0.19	-.69
					1.00	**0.81**	**0.79**	**0.77**	**0.81**	-.80	-.37	-.30	-.21	0.18	-.65
Visits in the past week (F)						1.00	**0.97**	**0.95**	**0.97**	-.79	-.56	-.56	-.44	-.05	-.74
						1.00	**0.97**	**0.94**	**0.97**	-.76	-.44	-.36	-.32	-.08	-.66
Mean visits in the past week per BA (G)							1.00	**0.98**	**0.98**	-.78	-.49	-.49	-.37	-.04	-.72
							1.00	**0.99**	**0.98**	-.71	-.34	-.27	-.27	-.07	-.61
% of BAs receiving ≥1 visit in past week (H)								1.00	**0.95**	-.77	-.50	-.50	-.41	-.02	-.71
								1.00	**0.95**	-.69	-.31	-.25	-.26	-.06	-.59
Standard deviation in visits among BAs in past week (I)									1.00	-.77	-.50	-.49	-.36	-.04	-.71
									1.00	-.72	-.36	-.29	-.27	-.07	-.62
Total visits (J)										1.00	**0.64**	**0.60**	**0.53**	-.22	0.60
										1.00	**0.48**	**0.38**	**0.38**	-.22	0.62
Mean visits per BA (K)											1.00	**0.83**	**0.78**	-.19	0.55
											1.00	**0.86**	**0.52**	-.12	0.44
% of BAs receiving ≥10 visits (L)												1.00	**0.61**	-.17	0.57
												1.00	**0.30**	-.10	0.35
Standard deviation in visits among BAs (M)													1.00	-.18	0.46
													1.00	-.08	0.35
Current coaching infidelity (N)														1.00	-.08
														1.00	-.13
Cumulative coaching infidelity (O)															1.00
															1.00

Table 2.6 Effect modification of coaching intensity by months since start of the intervention. All p-values were calculated using score tests. The p-value of the overall significance of coaching reflects the joint null hypothesis that both the main effect of coaching and the coaching-by-time interaction term are equal to zero.

	EBP Adherence N=2,083 births		**Primary Composite** Maternal morbidity, maternal or infant mortality (n/N=12,062/79,777)		**Secondary Composite** Maternal or infant mortality (n/N=3,907/80,234)	
Coaching Intensity	**p-value for interaction term**	**p-value for overall significance of coaching**	**p-value for interaction term**	**p-value for overall significance of coaching**	**p-value for interaction term**	**p-value for overall significance of coaching**
Coaching Frequency (past month)						
Visits in the past month	0.09	0.62	0.09	0.82	0.11	0.71
Mean visits in the past month per BA	0.08	0.28	0.09	0.94	0.11	0.25
% of BAs receiving ≥1 visit in past month	0.06	0.15	0.97	0.38	0.26	0.70
Standard deviation in visits among BAs in past month	0.07	0.53	0.32	0.48	0.44	0.49
Coaching Frequency (past week)						
Visits past week	0.07	0.43	0.02	0.06	0.02	0.14
Mean visits in the past week per BA	0.11	0.74	0.01	0.10	0.04	0.08
% of BAs receiving ≥1 visit in past week	0.09	0.64	0.04	0.28	0.05	0.34
Standard deviation in visits among BAs in past week	0.09	0.52	0.02	0.15	0.03	0.31
Cumulative coaching						
Total visits	0.23	0.09	0.12	0.22	0.09	0.77
Mean visits per BA	<.01	0.04	0.91	0.03	0.54	0.43
% of BAs receiving ≥10 visits	0.02	0.11	0.98	0.26	0.79	0.68
Standard deviation in visits among BAs	0.05	0.04	0.97	0.31	0.98	0.45
Coaching infidelity						
Current coaching infidelity	0.04	0.06	0.92	0.46	0.44	0.54
Cumulative coaching infidelity	0.63	0.14	0.07	0.53	0.85	0.39

Table 2.7 Effects of coaching frequency on EBP adherence using a one-week time horizon. Effects are reported for both a one-unit increase and for increasing each continuous coaching metric from its 25th percentile to its 75th percentile, or by one interquartile range (IQR). Results are from a generalized linear with an identity link. Standard errors are estimated using the empirical variance with an exchangeable working covariance structure to account for clustering at the facility level. (N=2,083 births)

		Model 1[1]			Model 2[2]		
Coaching Intensity	**Units in IQR increase**	**Δ in practices adhered to associated with one-unit increase (95% CI)**	**Δ in practices adhered to associated with IQR increase (95% CI)**	**p-value**	**Δ in practices adhered to associated with one-unit increase (95% CI)**	**Δ in practices adhered to associated with IQR increase (95% CI)**	**p-value**
Coaching Frequency							
Visits past week	1.0	0.7 (0.3, 1.0)	0.7 (0.3, 1.0)	<.01	0.3 (-0.1, 0.6)	0.3 (-0.1, 0.6)	0.14
Mean visits in the past week per BA	0.3	2.8 (1.5, 4.2)	0.8 (0.4, 1.2)	0.01	1.6 (0.2, 3.0)	0.5 (0.1, 0.9)	0.03
% of BAs receiving ≥1 visit in past week	30%	4.0 (2.0, 6.0)	1.2 (0.6, 1.8)	0.01	2.2 (0.3, 4.1)	0.7 (0.1, 1.2)	0.04
Standard deviation in visits among BAs in past week	0.5	2.0 (0.9, 3.0)	1.0 (0.5, 1.5)	0.01	0.8 (-0.3, 1.9)	0.4 (-0.2, 0.9)	0.14

[1]Adjusted for whether the facility was in a high-priority district, distance to district hospital, facility staff size, facility delivery load, whether birth occurred on the same day as a coaching visit

[2] Adjusted for everything in Model 1 plus months since start of the intervention (no non-linear effects of time detected).

Table 2.8 Effects of coaching frequency on health outcomes using a one-week time horizon**.** Effects are reported for increasing each continuous coaching metric from its 25th percentile to its 75th percentile, or by one interquartile range (IQR). Results are from a generalized linear model with a log link and binomial distribution. Standard errors are estimated using the empirical variance with an exchangeable working covariance structure.

		Primary Composite Maternal morbidity or maternal or infant mortality (n/N=12,062/79,777)				**Secondary Composite** Maternal or infant mortality (n/N=3,907/80,234)			
	Units in one	**Model 1[1]**		**Model 2[2]**		**Model 1[1]**		**Model 2[2]**	
Coaching Intensity	**IQR**	**RR (95% CI)**	**p-value**	**RR (95% CI)**	**p-value**	**RR (95% CI)**	**p-value**	**RR (95% CI)**	**p-value**
Coaching Frequency									
Visits past week	1.0	1.01 (0.99, 1.03)	0.41	0.99 (0.97, 1.02)	0.71	0.99 (0.95, 1.03)	0.69	1.00 (0.95, 1.06)	0.93
Mean visits in the past week per BA	0.3	1.01 (0.99, 1.04)	0.27	1.00 (0.98, 1.03)	0.82	0.99 (0.95, 1.03)	0.70	1.00 (0.96, 1.05)	0.97
% of BAs receiving ≥1 visit in past week	30%	1.02 (0.99, 1.06)	0.18	1.01 (0.98, 1.05)	0.40	1.01 (0.95, 1.07)	0.76	1.03 (0.97, 1.10)	0.30
Standard deviation in visits among BAs in past week	0.5	1.02 (0.99, 1.05)	0.21	1.01 (0.97, 1.05)	0.65	1.00 (0.95, 1.05)	0.99	1.03 (0.96, 1.10)	0.40

[1]Adjusted for whether the facility was in a high-priority district, distance to district hospital, facility staff size, facility delivery load, whether birth occurred on the same day as a coaching visit

[2] Adjusted for everything in Model 1 plus months since start of the intervention (no non-linear effects of time detected).

Chapter Three: Impact of PEPFAR's Prevention of Mother to Child Transmission of HIV program funding in Kenya: A quasi-experimental evaluation

ABSTRACT

Background: From 2004-2014, the President's Emergency Plan for AIDS Relief (PEPFAR) invested over US$248,000,000 in Prevention of Mother-to-Child Transmission (PMTCT) of HIV in Kenya. Concurrently, under-five mortality in Kenya decreased by half. The extent to which this decrease is attributable to PEPFAR is unknown.

Methods: Using 2004-2014 Country Operational Reports, we linked funding to Demographic and Health and AIDS Indicator Surveys. We estimated the impact of annual (ANN-PCF) and cumulative (CUM-PCF) per capita PEPFAR funding on infant and neonatal mortality and HIV testing at antenatal care (ANC), adjusting for year, province, and respondent characteristics using a quasi-experimental dose-response analysis.

Results: Among survey respondents, data was available for 26,876 infants and 20,775 mothers. A $0.33 increase in ANN-PCF, corresponding to the difference between the 75th and 25th (IQR) percentiles of funding, was significantly associated with a 16% (95% CI: 4-27%) reduction in infant mortality after a 1-year lag. This mortality reduction persisted after 2- and 3-year lags. A 1-IQR increase in CUM-PCF was associated with a 14% decrease in infant mortality (95% CI: 1-25%), which persisted after a 1-year lag, and with a 7% increase in HIV testing at ANC (95% CI: 3-11%), which intensified with subsequent lags. Funding was unassociated with neonatal mortality. Between 2004-2014, sustained funding levels of $0.33 ANN-PCF would have averted 118,039 to 273,924 infant deaths.

Conclusions: Evidence from publicly-available data suggests that PEPFAR's PMTCT funding reduced infant mortality and increased HIV testing at ANC in Kenya. The full impact of funding may not be felt until several years after allocation.

INTRODUCTION

Between 1988 and 2003, Kenya experienced a 32% increase in under-five mortality, which has been partially attributed to the HIV epidemic (121, 122). In response, Kenya established Prevention of Mother-to-Child Transmission of HIV (PMTCT) programs in over 10,000 facilities (123). The services provided by PMTCT programs are designed to prevent the vertical transmission of HIV during pregnancy, delivery, and breastfeeding and include HIV testing, post-test counseling for HIV-positive and -negative women, and access to anti-retroviral medications for HIV-positive women. Kenya's investment in PMTCT was supported by the U.S. President's Emergency Fund for AIDS Relief (PEPFAR), which contributed over $248 million to PMTCT programs in Kenya between 2004 and 2014 (124).

Although PEPFAR's investments in PMTCT coincided with a halving of Kenya's under-five mortality rate (125), it is unknown whether this improvement can be causally attributed to PEPFAR funding for PMTCT. PMTCT is critical to the survival of children born to HIV-positive mothers. Without PMTCT, 25% of children born to HIV-positive mothers become HIV-positive, and, without treatment, 50% of HIV-positive children in low-resource settings die before their second birthday (126, 127). However, child mortality decreased in most sub-Saharan African countries during the 2000s (128), and regional trends, rather than PEPFAR funding, could explain all or part of Kenya's reductions in child mortality. Additionally, PEPFAR-funded activities targeting adults could have displaced essential newborn and infant health services (129, 130), resulting in worsened child health outcomes.

Although PEPFAR focus countries have experienced greater reductions in adult mortality than non-focus countries (131, 132), previous studies have not found corresponding reductions in

child mortality (130, 133, 134). However, prior research has not specifically assessed the impact of PEPFAR funding for PMTCT, one of PEPFAR's activities most directly linked to children, on child health outcomes. In this paper, we use a quasi-experimental dose-response model to evaluate whether the magnitude of PEPFAR funding for PMTCT is casually associated with increased probability of receiving HIV counseling, testing, and test results as part of antenatal care (ANC) and reduced risk of neonatal and infant mortality in Kenya.

METHODS

Data Sources

Health outcome data came from the publicly available Kenya Demographic and Health Surveys (KDHS) and AIDS Indicator Surveys (KAIS), which use stratified two-stage clustered random sampling to select nationally-representative samples. Our analysis included data from the five most recent surveys: KDHS 2003, 2008/2009 and 2014 and KAIS 2007 and 2012 (123, 135-138). HIV testing at ANC was measured among female respondents who had given birth within five years of the 2007, 2008/2009, 2012 surveys and two years of the 2014 survey and is defined as receiving counseling on PMTCT, an HIV test, and test results during ANC. For women with multiple births, ANC data was gathered for the most recent birth. Neonatal and infant mortality, defined as death within the first month and first year of life, respectively, were assessed using female respondents' birth histories from the 2003, 2008/2009, and 2014 surveys. For each live birth, mothers reported their child's birth date, vital status, and death date, if applicable. Because maternal HIV is a common cause of maternal and infant death, birth histories can underestimate infant mortality in generalized HIV epidemics (139); however, bias can be reduced by considering recent births (140). Therefore, neonatal and infant mortality

were assessed among children born 1-60 and 12-60 months prior to the interview date. Although the 2012 KAIS gathered data on abbreviated birth histories, we excluded it in our primary analysis because neonatal and infant mortality was not included in the final report of the survey (see Supplementary Materials Methods Section A) (123). Data on the independent variable, PEPFAR funding for PMTCT, was extracted from publicly available Country Operational Plans (COPs), which describe annual planned expenditures (see Supplementary Materials Methods Section B) (141).

Statistical Methods

We assessed the association between per capita PEPFAR funding for PMTCT and health outcomes using a dose-response model, a type of quasi-experimental design used for causal inference (142). For annual per capita funding (ANN-PCF), each individual's dose equaled the amount of PEPFAR funding allocated to their province of residence in their year of birth (or, for HIV testing at ANC, in the year they gave birth) divided by the province's full population (143) .For cumulative per capita funding (CUM-PCF), each individual's dose was calculated using the cumulative total of PEPFAR funding for PMTCT allocated to their province from the beginning of PEPFAR until their year of birth. We investigated zero to three-year lags between funding allocation and the time that funding was hypothesized to have effects (see Supplementary Materials Table 3.5). Individuals without funding data were excluded from primary analyses but included in sensitivity analyses.

We estimated risk ratios and 95% confidence intervals associated with PEPFAR funding for PMTCT using weighted generalized estimating equations (GEEs) that created a representative sample across surveys and stepwise selection of restricted cubic splines that assessed for

potential non-linear relationships between funding and health outcomes (23, 24) All models adjusted for province and controlled for calendar year using restricted cubic splines. Fully-adjusted models further controlled for household wealth quintile, water and sanitation access, urban/rural status, and mosquito net ownership; respondent education, ethnicity, religion, marital status, age, parity, and exposure to mass media; and, for neonatal and infant mortality, child's sex, preceding birth interval, and birth order rather than parity (see Supplementary Materials Methods Section C for details). To estimate the number of lives saved by PEPFAR funding for PMTCT, we used the one-year lagged model to predict the number of infants who would have died between 2004 and 2014 under different levels of ANN-PCF (see Supplementary Materials Methods Section D for details).

We evaluated the sensitivity of our results to missing exposure data by conducting analyses assuming that province-years with missing funding data received \$0.04, \$0.23, \$0.32, \$0.56, and \$0.93 in ANN-PCF (reflecting the minimum, 25th percentile, median, 75th percentile, and maximum of observed funding levels, respectively) and, for infant mortality, using inverse probability weighting (see Supplementary Materials Methods Section E) (144). We also assessed whether the effect of PEPFAR funding on infant mortality differed by maternal HIV status among the subset of infants with known maternal HIV status based on HIV testing conducted in the 2003 and 2008 surveys (see Supplementary Materials Methods Section F for details).

RESULTS

Infant Mortality

The 2003, 2008/09, and 2014 surveys included birth histories for 128,199 children, 26,876 of whom were born 12-60 months before the survey. Of these, 1,222 died within the first year of life. After a one-year lag, a $0.33 increase in ANN-PCF, corresponding to the difference between the 75th and 25th percentiles (IQR) of observed ANN-PCF data, was associated with a significant 16% reduction in infant mortality (95% CI: 4-27) in the fully adjusted model (Table 3.1). This reduction was sustained after two- and three-year lags (Figure 3.1) but was not significant in the unlagged model. For CUM-PCF, a $0.83 increase (reflecting the IQR of CUM-PCF) was similarly associated with a 14% decrease in infant mortality (95% CI: 1-25%) in the unlagged model. This reduction was sustained after a 1-year lag, with associations attenuating and becoming non-significant after subsequent lags.

Table 3.1 Risk ratios and 95% confidence intervals for infant mortality.

		ANN-PCF		CUM-PCF
		Minimally Adjusted[1]	**Fully Adjusted[2]**	**Fully Adjusted[2]**
No lag	RR for $0 to 1 IQR increase[3]	0.88 (0.73, 1.07)	0.90 (0.75, 1.07)	0.86 (0.75, 0.99)
	RR for $0 to maximum increase[4]	0.70 (0.41, 1.20)	0.73 (0.44, 1.21)	0.37 (0.15, 0.91)
	p-value	0.20	0.22	0.04
1-year lag	RR for $0 to 1 IQR increase[3]	0.83 (0.72, 0.95)	0.84 (0.73, 0.96)	0.86 (0.74, 0.99)
	RR for $0 to maximum increase[4]	0.59 (0.40, 0.87)	0.61 (0.42, 0.90)	0.36 (0.14, 0.91)
	p-value	0.01	0.01	0.04
2-year lag	RR for $0 to 1 IQR increase[3]	1.19 (0.82, 1.74)	0.86 (0.75, 0.99)	0.93 (0.81, 1.06)
	RR for $0 to maximum increase[4]	0.51 (0.32, 0.81)	0.66 (0.44, 0.98)	0.60 (0.24, 1.47)
	p-value	0.26[5]	0.04	0.27
3-year lag	RR for $0 to 1 IQR increase[3]	0.83 (0.71, 0.97)	0.84 (0.72, 0.98)	0.91 (0.78, 1.06)
	RR for $0 to maximum increase[4]	0.60 (0.39, 0.92)	0.61 (0.40, 0.95)	0.55 (0.20, 1.50)
	p-value	0.02	0.03	0.24

[1]Adjusted only for province and calendar year.
[2]Adjusted for province, calendar year, household wealth quintile, water and sanitation access, urban/rural status and mosquito net ownership, maternal age at birth, education, ethnicity, religion, marital status, and exposure to mass media; and child's sex, short preceding birth interval, and birth order.
[3]A $0 to 1 IQR increase in PEPFAR funding for PMTCT corresponds to a $0.33 change in annual per capita funding (ANN-PCF) and a 0.83 change in cumulative per capita funding (CUM-PCF).
[4]A $0 to maximum increase in PEPFAR funding for PMTCT corresponds to a $0.93 change in ANN-PCF and a $5.46 change in CUM-PCF.

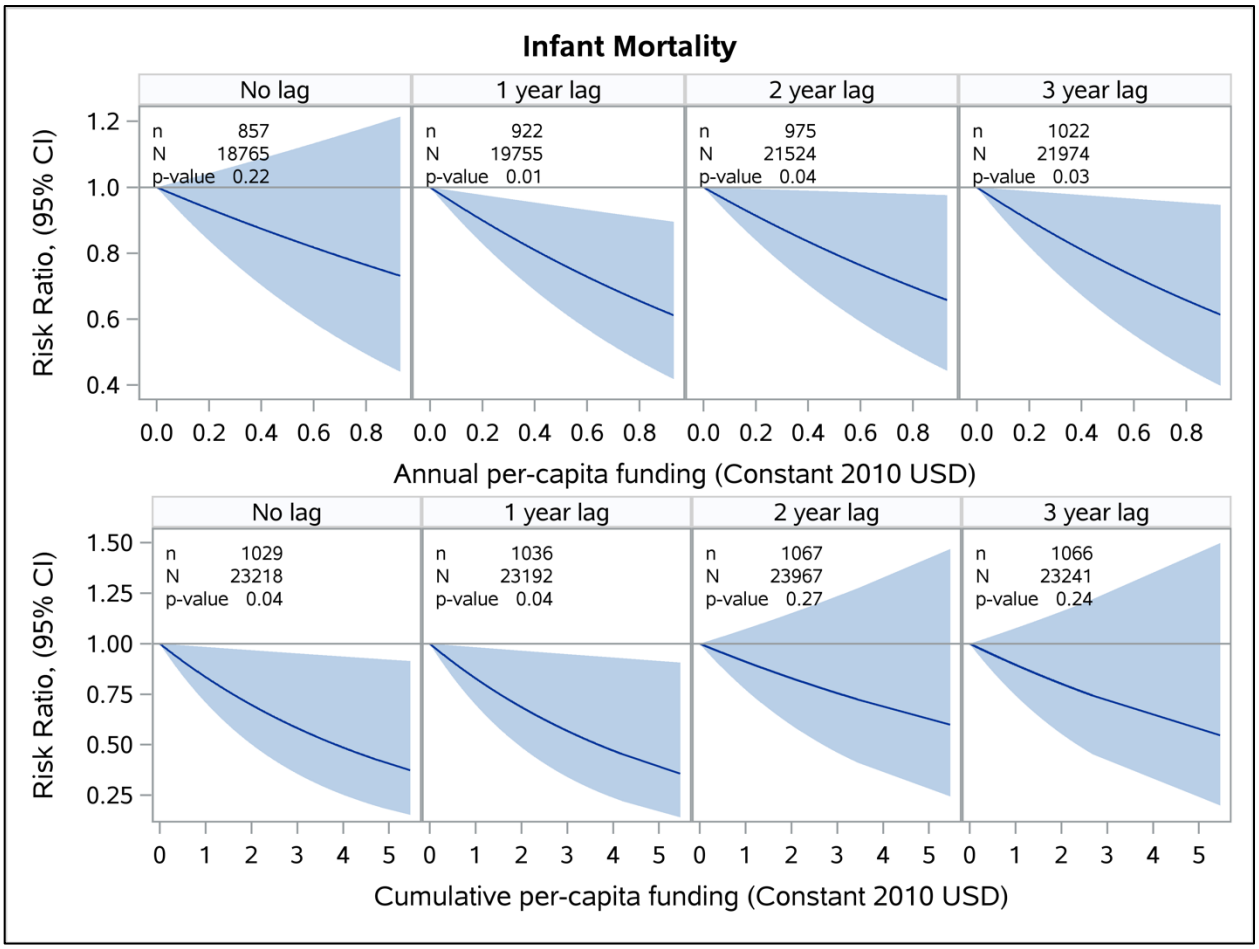

Figure 3.1 Dose-response relationship between per capita PEPFAR funding for PMTCT and infant mortality. Each panel presents the number of deaths (n) out of the number of infants with complete exposure data (N). (Estimates are adjusted for province; year of birth; household wealth quintile, water and sanitation access, urban/rural status and mosquito net ownership; maternal age at birth, education, ethnicity, religion, marital status, and exposure to mass media; and child's sex, birth order, and short preceding birth interval.)

HIV Testing at ANC

The 2007, 2008-2009, 2012, and 2014 surveys included data on 57,721 female respondents, 21,488 of whom had recently given birth. Of the 20,775 women with data on HIV testing, 11,984 received HIV testing during ANC. ANN-PCF was not associated with HIV testing at ANC in any model (Table 3.2). However, after a one-year lag, a $0.83 increase in CUM-PCF was associated with a 7% increase in HIV testing at ANC (95% CI: 4-15%). This association intensified with subsequent lags: a $0.83 increase in CUM-PCF was associated with a 9% increase after a two-year lag (95% CI: 5-14%) and a 19% increase after a three-year lag (95% CI: 11-28%, Figure 3.2).

Table 3.2 Risk ratios and 95% confidence intervals for HIV testing at ANC.

		ANN-PCF		CUM-PCF
		Minimally Adjusted[1]	**Fully Adjusted[2]**	**Fully Adjusted[2]**
No lag	RR for $0 to 1 IQR increase[3]	0.99 (0.96, 1.03)	0.98 (0.94, 1.01)	0.96 (0.89, 1.04)
	RR for $0 to maximum increase[4]	0.97 (0.88, 1.08)	0.93 (0.84, 1.04)	1.10 (0.85, 1.42)
	p-value	0.61	0.20	0.33[5]
1-year lag	RR for $0 to 1 IQR increase[3]	1.01 (0.97, 1.06)	1.01 (0.97, 1.05)	1.07 (1.03, 1.11)
	RR for $0 to maximum increase[4]	1.04 (0.93, 1.16)	1.03 (0.92, 1.16)	1.54 (1.20, 1.98)
	p-value	0.48	0.57	0.0008
2-year lag	RR for $0 to 1 IQR increase[3]	1.01 (0.97, 1.04)	1.02 (0.98, 1.06)	1.09 (1.05, 1.14)
	RR for $0 to maximum increase[4]	1.02 (0.93, 1.12)	1.05 (0.95, 1.17)	1.76 (1.35, 2.31)
	p-value	0.67	0.33	<0.0001
3-year lag	RR for $0 to 1 IQR increase[3]	1.00 (0.97, 1.03)	1.02 (0.99, 1.05)	1.19 (1.11, 1.28)
	RR for $0 to maximum increase[4]	1.01 (0.93, 1.09)	1.06 (0.98, 1.16)	1.65 (1.20, 2.28)
	p-value	0.88	0.15	<0.0001[5]

[1]Adjusted only for province and calendar year.
[2]Adjusted for province, calendar year, household wealth quintile, water and sanitation access, urban/rural status and mosquito net ownership and maternal age at last birth, education, ethnicity, religion, marital status, parity and exposure to mass media.
[3]A $0 to 1 IQR increase in PEPFAR funding for PMTCT corresponds to a $0.33 change in annual per capita funding (ANN-PCF) and a 0.83 change in cumulative per capita funding (CUM-PCF).
[4]A $0 to maximum increase in PEPFAR funding for PMTCT corresponds to a $0.93 change in ANN-PCF and a $5.46 change in CUM-PCF.
[5]A significant departure from linearity was detected, the p-value reflects the significance of the non-linear relationship.

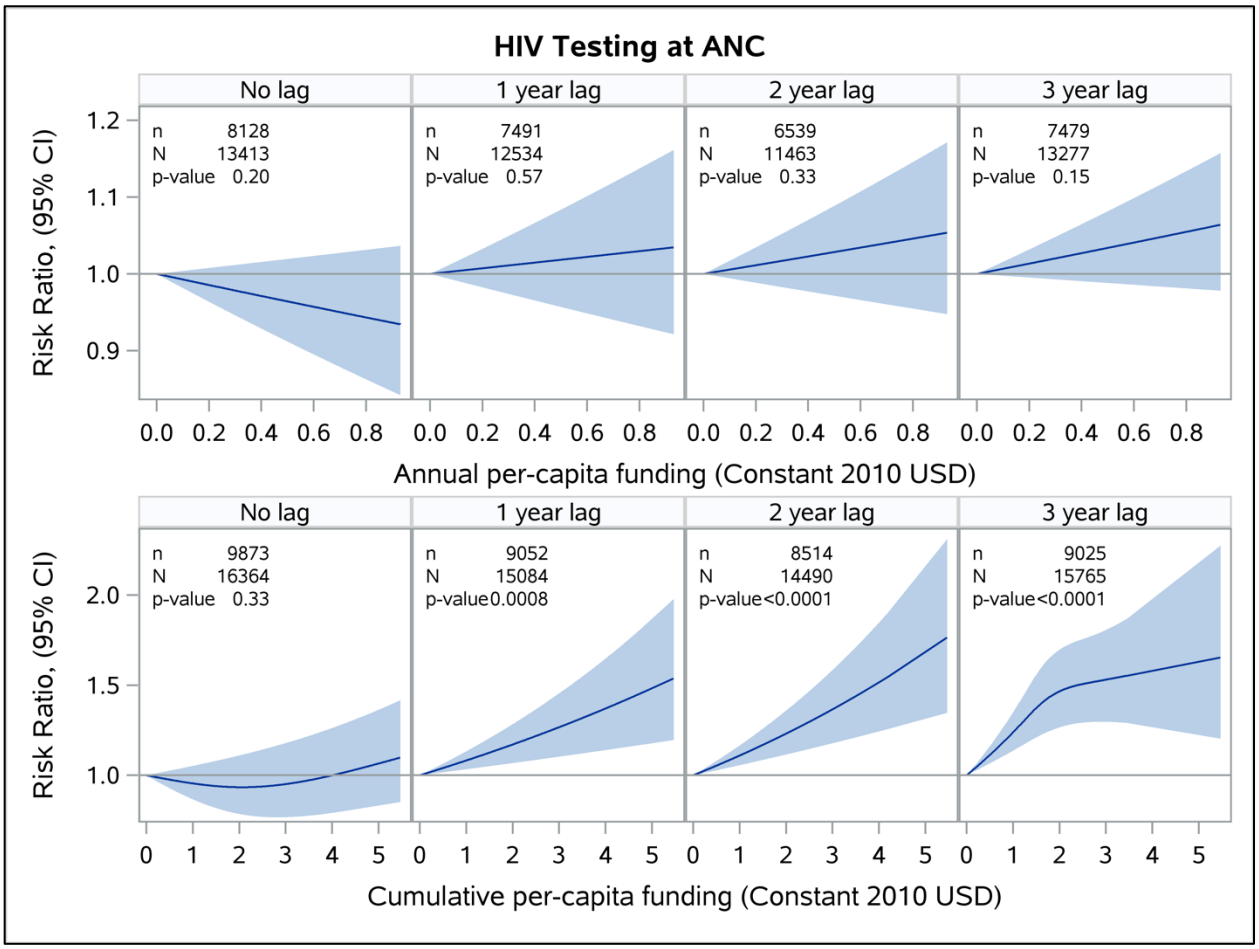

Figure 3.2 Dose-response relationship between per capita PEPFAR funding for PMTCT and HIV Testing at ANC, defined as receiving counseling on PMTCT, being tested for HIV, and receiving the results of this HIV test during an antenatal care. Each panel presents the number of women tested (n) out of the number of women reporting a recent birth with complete exposure data (N). (Estimates are adjusted for province; year of birth; household wealth quintile, water and sanitation access, urban/rural status and mosquito net ownership and maternal age at last birth, education, ethnicity, religion, marital status, parity, and exposure to mass media.)

Neonatal Mortality

Our analysis included 33,181 neonates born 1-60 months before the survey, 822 of whom died during the first month of life. Neonatal mortality was not associated with ANN-PCF or CUM-PCF (Table 3.3, Supplementary Materials Figure 3.4).

Table 3.3 Risk ratios and 95% confidence intervals for neonatal mortality.

		ANN-PCF		CUM-PCF
		Minimally Adjusted[1]	**Fully Adjusted[2]**	**Fully Adjusted[2]**
No lag	RR for $0 to 1 IQR increase[3]	1.17 (0.92, 1.49)	1.21 (0.96, 1.52)	0.98 (0.83, 1.17)
	RR for $0 to maximum increase[4]	1.56 (0.79, 3.06)	1.71 (0.90, 3.27)	0.90 (0.29, 2.78)
	p-value	0.22	0.12	0.85
1-year lag	RR for $0 to 1 IQR increase[3]	1.03 (0.83, 1.28)	1.05 (0.84, 1.31)	1.03 (0.86, 1.23)
	RR for $0 to maximum increase[4]	1.08 (0.59, 1.99)	1.15 (0.62, 2.14)	1.18 (0.36, 3.80)
	p-value	0.80	0.66	0.79
2-year lag	RR for $0 to 1 IQR increase[3]	0.99 (0.82, 1.20)	1.00 (0.83, 1.21)	1.11 (0.92, 1.34)
	RR for $0 to maximum increase[4]	0.98 (0.57, 1.68)	1.01 (0.60, 1.70)	2.04 (0.60, 6.96)
	p-value	0.94	0.98	0.26
3-year lag	RR for $0 to 1 IQR increase[3]	0.94 (0.75, 1.18)	0.96 (0.77, 1.19)	1.06 (0.86, 1.31)
	RR for $0 to maximum increase[4]	0.85 (0.45, 1.59)	0.88 (0.47, 1.65)	1.47 (0.36, 5.98)
	p-value	0.60	0.69	0.59

[1]Adjusted only for province and calendar year.
[2]Adjusted for province, calendar year, household wealth quintile, water and sanitation access, urban/rural status and mosquito net ownership, maternal age at birth, education, ethnicity, religion, marital status, and exposure to mass media; and child's sex, short preceding birth interval, and birth order.
[3]A $0 to 1 IQR increase in PEPFAR funding for PMTCT corresponds to a $0.33 change in annual per capita funding (ANN-PCF) and a 0.83 change in cumulative per capita funding (CUM-PCF).
[4] A $0 to maximum increase in PEPFAR funding for PMTCT corresponds to a $0.93 change in ANN-PCF and a $5.46 change in CUM-PCF.

Lives Saved

Based on the one-year lagged model, increasing PEPFAR funding for PMTCT from $0 to $0.33 in ANN-PCF would have averted 118,039 infant deaths between 2004-2014 and increasing funding from $0 to $0.93 in ANN-PCF would have averted 286,438 infant deaths over the same period. These estimates are conservative and do not include benefits accrued among mothers or among children older than one. When KAIS 2012 was included, the estimated lives saved was

appreciably larger: a $0.33 increase in ANN-PCF was associated with a 36% reduction in infant mortality after a one-year lag (95% CI: 12-53%), which, over the 2004-2014 period, would correspond to 273,924 infant deaths averted at funding levels of $0.33 in ANN-PCF and 547,179 infant deaths averted at funding levels of $0.93.

Sensitivity Analyses

Sensitivity analyses for our unlagged, one-year lagged, and three-year lagged models generally corresponded to the main results. For infant mortality, ANN-PCF remained associated with a 13-19% reduction when province-years with missing data were assigned $0.04-0.32 in ANN-PCF or with inverse probability weighting (Supplemental Materials Tables 3.6-3.8). When using a two-year lag, discrepancies reflected the detection of new non-linear results. When we investigated the effect of PEPFAR funding by maternal HIV status, PEPFAR was significantly protective against mortality among infants born to HIV-negative mothers in the unlagged and one-year lagged model but not among infants born to HIV-positive mothers (Supplementary Materials 3.9). However, due to the small number of deaths among infants with known maternal HIV status (214 deaths among the positive and 45 among the negatives), there was no evidence that PEPFAR funding had significantly different effects in the two groups.

DISCUSSION

Our study joins a growing body of literature that suggests PEPFAR has benefited population health (131-133, 145-147). Using publicly available data, we found evidence that PEPFAR funding for PMTCT is causally associated with reduced infant mortality and increased HIV testing at ANC in Kenya. Our quasi-experimental dose-response design provides stronger evidence for a causal effect of PEPFAR funding than other designs, such as pre-post designs,

commonly used to assess PEPFAR's effectiveness, because it directly adjusts for secular time trends (148).

Earlier studies had not found significantly greater reductions in infant mortality among PEPFAR focus countries relative to non-focus countries; however, they did report non-significant protective effects (133, 134). Our significant findings may stem from one or more factors: funding for PMTCT may impact infant mortality more directly than overall PEPFAR funding; our study period (2004-2014) was longer than for previous studies (2004-2010); our study was not confounded by country-level factors; and Kenya received more PEPFAR funding than most other PEFPAR countries and may have experienced correspondingly greater effects. Despite its positive impact on infant mortality and HIV testing at ANC, PEPFAR funding was unassociated with neonatal mortality, perhaps because many pediatric HIV-infections occur after the neonatal period (127) or because neonatal mortality may be more resistant to public health interventions than post-neonatal mortality (128).

Delayed Effects of Annual Funding

Annual funding was unassociated with reduced infant mortality in the year of allocation but became beneficial at later lags (Figure 3.1). Although non-significant, similar trajectories were observed for HIV testing at ANC and neonatal mortality, suggesting that the full impact of annual funding may not be observable for several years. This delay may reflect logistical delays as PEPFAR funds are absorbed by local PMTCT programs but may also reflect biologic realities: there is a 9-month lag between conception and birth and another 12-month lag between birth and ascertaining infant survival. Both logistic and biologic factors should be considered when defining a program evaluation's time horizon.

Cumulative Funding and Threshold Effects

When using cumulative, rather than annual, funding we observed more significant associations after shorter lags. Compared to annual funding, cumulative funding may better reflect sustained investments in infrastructure and personnel. The significant non-linear association between 3-year lagged cumulative funding and HIV-testing at ANC may suggest a threshold effect for cumulative per capita funding. However, as of 2014 only one (Nyanza) of eight provinces had received >$2.68 in CUM-PCF, this apparent threshold may be largely driven by a single province and we did not observe significant thresholds in any of the other adjusted models. The joint findings of the ANN-PCF and CUM-PCF models suggest that there is a delayed effect of annual funding on PMTCT-related outcomes, that the effect of cumulative funding is greater than the effect of spending in any individual year, and that there is no evidence of diminishing returns to PEPFAR's investment in PMTCT programs.

Robustness of results

Our findings were relatively robust to sensitivity analyses assessing possible bias due to missing funding data. Only under unlikely scenarios where missing province-years received very high ANN-PCF ($0.56-0.93) did our results for infant mortality become attenuated across all lags. Discrepancies between the main analysis and sensitivity analyses either reflected the detection of new non-linear relationships or occurred under unlikely scenarios in which missing province-years received very high ANN-PCF. These analyses suggest that, despite limitations of COPs as a source of regional funding data, our findings are unlikely to be explained by missing data patterns.

When we examined whether the effect of funding on infant mortality differed by mothers' HIV status, PEPFAR was not associated with significantly lower infant mortality among infants of HIV positive mothers; however, power was severely limited by the small sample of mothers with known HIV status. In contrast, infants of HIV-negative mothers experienced reduced mortality, possibly reflecting positive spillover effects including improved health literacy, health professional training, and health system resources.

Limitations

We sought to minimize unmeasured confounding by adjusting for fixed effects of province, which controls for unobserved time-invariant province characteristics, and for year, which controls for nationwide secular trends. However, our analysis could be confounded by province characteristics that both vary over time and are correlated with PEPFAR funding for PMTCT, such as funding for anti-malarial campaigns. We strove to partially capture time-varying province characteristics by adjusting for respondent-level characteristics, such as mosquito net ownership. However, bias from time-varying confounding is possible, particularly if PEPFAR preferentially allocated funds to implementing partners working in provinces with improving health outcomes. Additionally, the data source for our exposure, COPs, describe annual planned expenditures, which may not reflect actual expenditures, and disaggregate funding by implementing partner rather than geography, limiting our ability to assign funding to provinces. We expect these administrative processes to result in random measurement error and underestimation of effects. However, systematic bias could occur if, for example, implementing partners working in provinces with strong health systems reported higher quality information in COPs than those working in provinces with weak health systems. Furthermore, our analysis

links province-level funding to individual-level outcomes and does not account for variation in the distribution of PEPFAR-funded activities within a province or variation in individuals' engagement with PEPFAR-funded PMTCT programs. Thus, these findings are best interpreted as the effect of living in a province receiving a given level of PEPFAR funding and likely underestimate the benefits of interacting directly with a PEPFAR-funded PMTCT program. Finally, because health outcome data was unavailable after 2014, our analysis does not capture effects of recent PEPFAR-funded PMTCT-related activities.

Public Health Relevance

This study illustrates how cost-effective, large-scale program evaluations can be conducted using pre-existing data sources. We relied exclusively on COPs, which are produced annually by all PEPFAR focus countries, and the Demographic and Health (DHS) and AIDS Indicator Surveys (AIS). At least one DHS or AIS has been conducted in all 31 countries producing COPs. Consequently, our approach may be applied in other geographic settings and adapted for other HIV-related health outcomes. PEPFAR and other international donors seeking to evaluate programmatic impact might consider routinely collecting annual data on financial expenditures disaggregated by geography. Linking this financial data to DHS, AIS, or ongoing PEPFAR-funded population-based HIV impact assessment (PHIA) surveys would enable future impact evaluations. Recently released subnational-level data on PEPFAR's programmatic activities for 2015-2016 (124) could also be used in future dose-response evaluations.

Investigating a dose-response between funding and health outcomes addresses policy questions about the effectiveness of programs and can inform allocation of future funds. Our study follows three previous studies investigating the magnitude of PEPFAR funding as an

exposure and resulting effect on intended health and behavioral outcomes, with varying findings (145, 146, 149). For example, Lo and colleagues did not observe an association between funding for abstinence and being faithful activities and reduced high-risk sexual behavior (149). Access to range of literature on the effects of funding for specific HIV prevention approaches can help policy makers decide which evidence-informed activities to fund and increase allocation of funds to approaches demonstrated to be effective.

CONCLUSION

PEPFAR funding for PMTCT was associated with substantially reduced infant mortality and increased HIV testing at ANC in Kenya. This research illustrates how pre-existing data sources can be used to conduct cost-effective, large-scale program evaluations in a robust and timely manner.

SUPPLEMENTAL MATERIALS

Methods Section A: Birth history data from KAIS 2012

Although the KAIS 2012 gathered data on abbreviated birth histories, we excluded it in our primary analysis of neonatal and infant mortality. This decision was driven by several factors. First, the KAIS 2012 did not report on birth history data in their final report, which may reflect the fact that women's birth histories, which have traditionally been a major focus of Demographic and Health Surveys such as the KDHS, have not traditionally been a major focus of AIDS Indicator Surveys, such as the KAIS. This omission may also reflect underlying concerns about the data quality. As shown in the table below, the five-year neonatal and infant mortality rate calculated using data from the KAIS 2012 was substantially lower than that calculated from other surveys, even when the time frame covered by those other surveys overlapped substantially with the time frame covered by the KAIS 2012.

Table 3.4 Five-year infant and neonatal mortality by survey year

		Infant mortality			Neonatal Mortality		
Survey	**Years Covered**	**Deaths**	**Infants**	**Mortality rate % , (95% CI)**	**Deaths**	**Neonates**	**Mortality rate % , (95% CI)**
KDHS 2003	1998-2003	349	4750	7.7, (6.7, 8.7)	197	5960	3.3, (2.7, 3.9)
KDHS 2008	2003-2009	254	4850	5.4, (4.3, 6.4)	176	6083	3.1, (2.4, 3.8)
KAIS 2012	2008-2013	52	3548	2.4, (0, 4.7)	30	4435	0.6, (0.3, 0.8)
KDHS 2014	2009-2014	619	17276	3.8, (3.4, 4.2)	449	21138	2.2, (1.9, 2.5)

Due to these underlying concerns, we choose to omit KAIS 2012 from our primary analysis of infant and neonatal mortality.

Methods Section B: Data Extraction from Country Operational Plans (COPs)

Each year, PEPFAR focus countries submit country operational plans (COPs) containing PEPFAR's annual planned expenditures disaggregated by implementation partner and budget code. We identified annual PMTCT funding from the COPs using the "MTCT" budget code. Narrative information from the COPs were used to assign dollar amounts to one of Kenya's eight provinces. Between 2004-2006, all financial data was redacted from the COPs. Overall, we assigned 53% of total planned PEPFAR expenditures for PMTCT to specific provinces; the remainder could not be assigned to specific provinces either because it had been allocated to nationwide programs (16%) or implementing partners working across province borders (10%) or because there was insufficient information to assign funding to a specific implementing partner (20%) (Figure 3.3). Dollar amounts were converted to 2010 USD using a GDP deflator. We set funding for pre-PEPFAR years to zero.

Methods Section C: Modeling the risk ratio

We estimated risk ratios and 95% confidence intervals associated with PEPFAR funding for PMTCT using generalized weighted estimating equations (GEEs), using exchangeable working covariance structures with robust standard errors to account for correlations within sampling clusters (23). To create a representative sample across surveys, we used denormalized weights calculated by multiplying the survey sampling weights by the number of individuals in the survey's sampling frame divided by the survey's sample size (150). To maximize statistical efficiency, we used the working binomial variance or, if models failed to converge, the working Poisson variance (108, 151). We assessed potential non-linear relationships between funding and health outcomes using restricted cubic splines (24), selecting spline terms using a SAS

macro available on the last author's website (111) and presenting the model with the linear term if non-linear relationships were non-significant. All models adjusted for province and controlled for calendar year using restricted cubic splines. Fully-adjusted models further adjusted for household characteristics, including household wealth quintile, water and sanitation access, urban/rural status, and mosquito net ownership, and respondent characteristics, including education level, ethnicity, religion, marital status, maternal age categorized as <20, 20-34, and ≥35 years of age, and exposure to mass media. For the outcomes of neonatal and infant mortality, fully adjusted models also adjusted for child's sex, short preceding birth interval (≤24 months) and birth order (rather than parity). To account for secular changes in wealth, we developed harmonized wealth quintiles using the first principal component of household assets measured across all surveys (152).

Methods Section D: Estimating the number of lives saved

To estimate the number of lives saved by PEPFAR funding for PMTCT, we used the one-year lagged model to predict the probability of infant mortality in each province-year assuming $0, $0.33, and $0.93 in ANN-PCF. We multiplied these predicted probabilities of death by the number of infants born in each province-year between 2004 and 2014 calculated by allocating the UN Population Division's national birth estimates proportionally to the World Bank Group's Subnational Population Database's adult population estimates (143).

Methods Section E: Inverse probability weighting for missing exposure data

For infant mortality, we used inverse probability weighting to adjust for possible bias due to the exclusion of infants with missing data on ANN-PCF (144). We predicted the probability that each infant was missing data on ANN-PCF using a logistic regression model that included the

outcome, with model covariates and two-way interactions being added to the model when significant at $p<0.05$. The inverse of this predicted probability, truncated at the 99th percentile, was used as a weight in the regression of infant mortality on ANN-PCF.

Methods Section F: Effect of PEPFAR funding for PMTCT by maternal HIV status

The sample size for our analysis stratified by maternal HIV status was limited because only the KDHS 2003 and KDHS 2008/2009 gathered both full birth history data and dried blot spots used to ascertain maternal HIV status. The sample size for this analysis consisted of 3,803 infants of HIV-negative mothers, 214 of whom had experienced infant mortality, and 369 infants of HIV-positive mothers, 45 of whom had experienced infant mortality. Because financial data from the COPs was retracted between 2004-2006, we were unable to fit dose-response models in the subset of infants whose mothers' HIV status was known. Instead, we modeled the zero- to three-year lagged effect of dichotomized pre- vs. post- PEPFAR funding, adjusted for year and province.

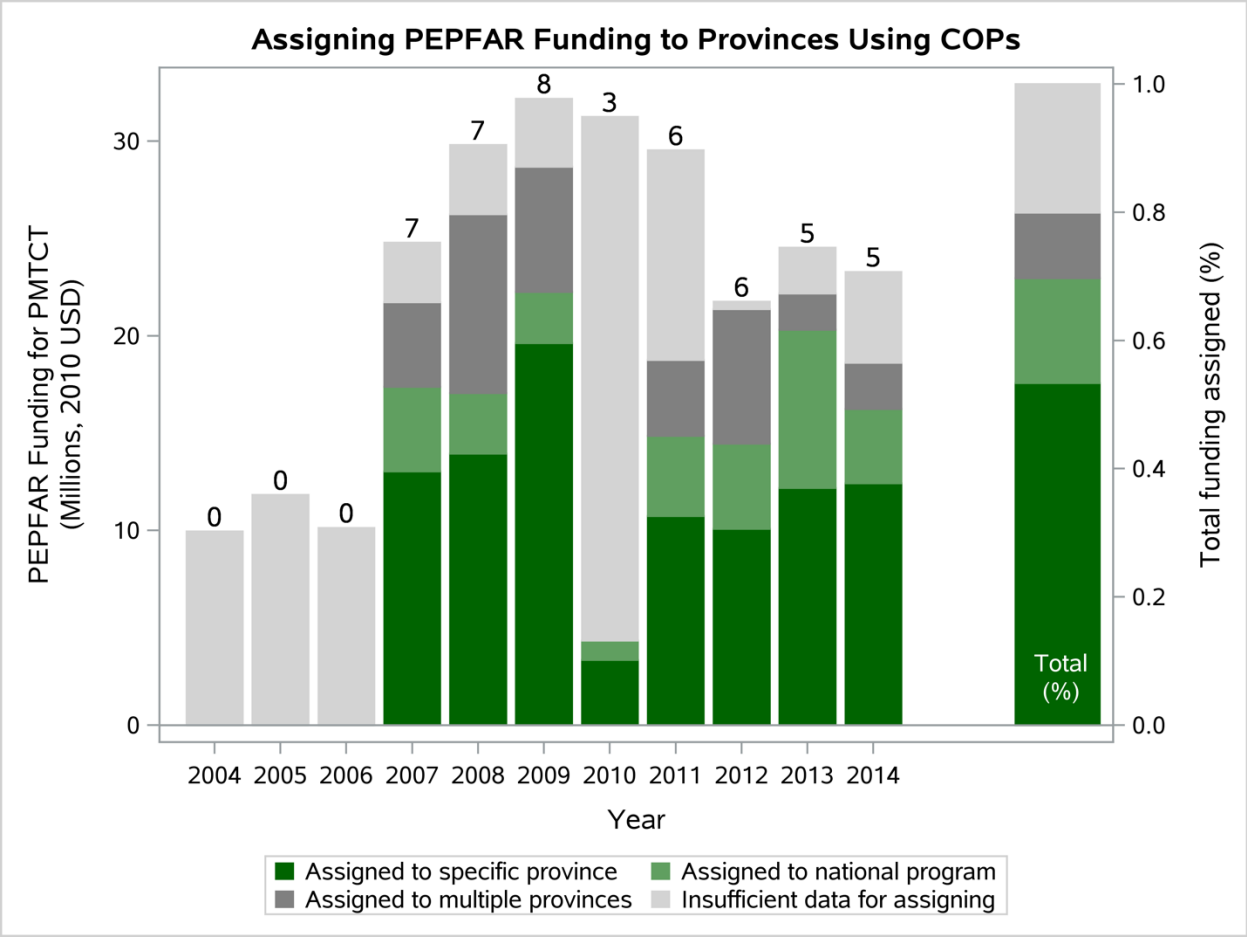

Figure 3.3 Assigning PEPFAR funding to Kenyan provinces using Country Operational Plans (COPs). The number at the top of each bar gives the number of provinces (out of 8) with data on PEPFAR funding for PMTCT in that year.

Table 3.5 Exemplary calculations for lagged annual and cumulative per-capita funding for a single province. "." designates a missing value.

			Annual per-capita funding for PMTCT				Cumulative per-capita funding for PMTCT			
Year	Funding for PMTCT (Millions, 2010 USD)	Population (Millions)	Unlagged	1-year lag	2-year lag	3-year lag	Unlagged	1-year lag	2-year lag	3-year lag
2000	$0	7.61	$0	$0	$0	$0	$0	$0	$0	$0
2001	$0	7.86	$0	$0	$0	$0	$0	$0	$0	$0
2002	$0	8.12	$0	$0	$0	$0	$0	$0	$0	$0
2003	$0	8.39	$0	$0	$0	$0	$0	$0	$0	$0
2004	.	8.67	.	$0	$0	$0	.	$0	$0	$0
2005	.	8.95	.	.	$0	$0	.	.	$0	$0
2006	.	9.24	.	.	.	$0	.	.	.	$0
2007	$2.57	9.54	$0.28	.	.	.	$0.28	.	.	.
2008	$3.08	9.85	$0.32	$0.28	.	.	$0.60	$0.28	.	.
2009	$5.60	10.17	$0.56	$0.32	$0.28	.	$1.16	$0.60	$0.28	.
2010	$2.03	10.50	$0.19	$0.56	$0.32	$0.28	$1.35	$1.16	$0.60	$0.28
2011	$1.97	10.83	$0.18	$0.19	$0.56	$0.32	$1.53	$1.35	$1.16	$0.60
2012	$0.52	11.17	$0.04	$0.18	$0.19	$0.56	$1.57	$1.53	$1.35	$1.16
2013	$3.02	11.52	$0.25	$0.04	$0.18	$0.19	$1.82	$1.57	$1.53	$1.35
2014	$2.54	11.88	$0.20	$0.25	$0.04	$0.18	$2.02	$1.82	$1.57	$1.53

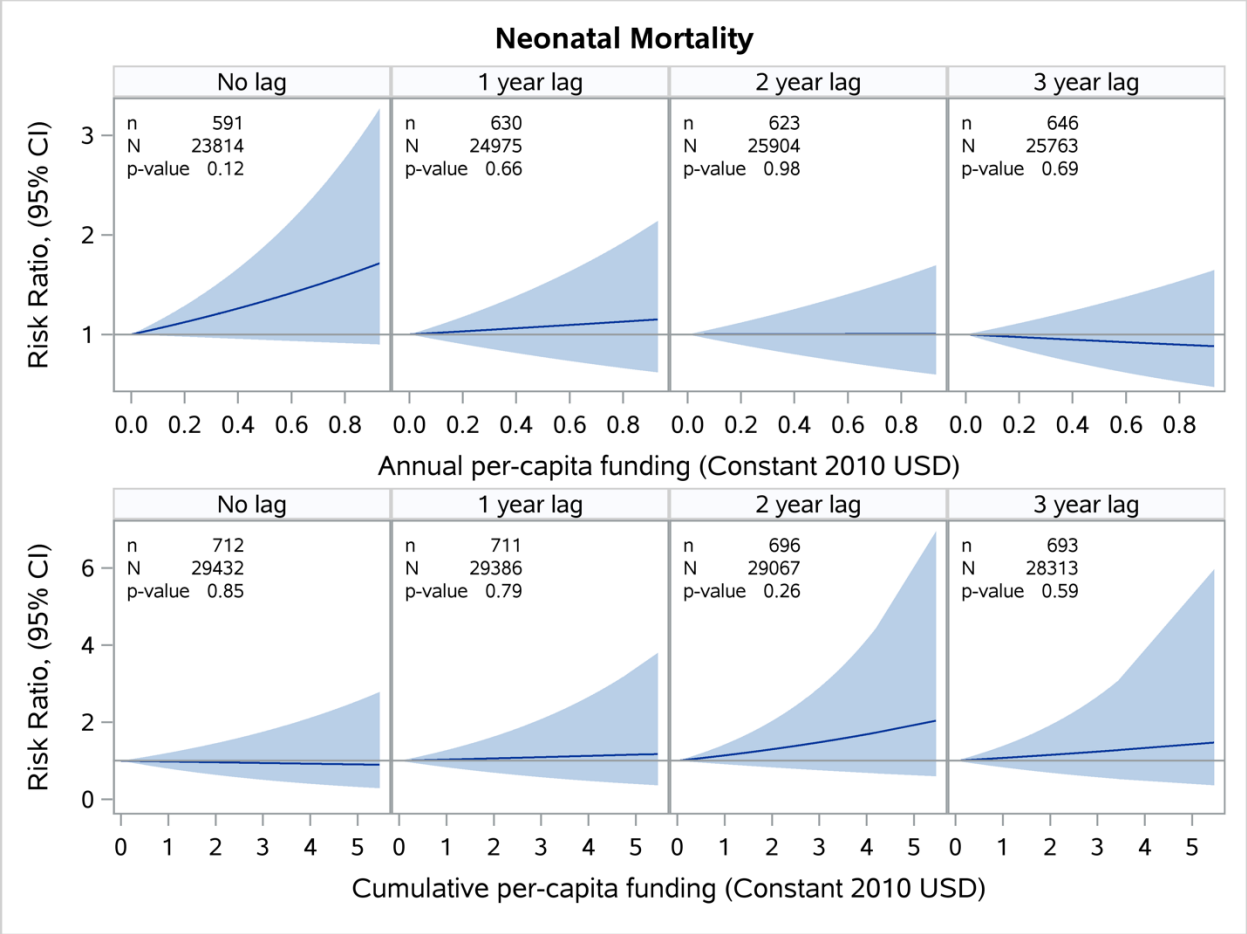

Figure 3.4 Dose-response relationship between per capita PEPFAR funding for PMTCT and neonatal mortality. Each panel presents the number of deaths (n) out of the number of infants with complete exposure data (N). Estimates are adjusted for province; year of birth; household wealth quintile, water and sanitation access, urban/rural status and mosquito net ownership; maternal age at birth, education, ethnicity, religion, marital status, and exposure to mass media; and child's sex, birth order, and short preceding birth interval.

Table 3.6 Summary of sensitivity analyses results for infant mortality. Cells include the risk ratio (RR) corresponding to a $0.33 increase in annual PEPFAR funding per capita. Table shows p-values for test of linear trend (L) or, if significant departures from linearity were detected, the test for non-linearity (NL) and test for the significance of the curve (C).

	Infant Mortality			
	No lag	**1 year lag**	**2 year lag**	**3 year lag**
U	0.88 (0.73, 1.07) L: 0.20	0.83 (0.72, 0.95) L: 0.01	1.19 (0.82, 1.74) NL: 0.04 C: 0.26	0.83 (0.71, 0.97) L: 0.02
A	0.90 (0.75, 1.07) L: 0.22	0.84 (0.73, 0.96) L: 0.01	0.86 (0.75, 0.99) L: 0.04	0.84 (0.72, 0.98) L: 0.03
S1	0.93 (0.80, 1.08) L: 0.36	0.91 (0.81, 1.02) L: 0.09	1.13 (0.83, 1.52) NL: 0.05 C: 0.31	0.87 (0.77, 0.98) L: 0.03
S2	0.89 (0.75, 1.06) L: 0.19	0.86 (0.75, 0.98) L: 0.02	1.18 (0.88, 1.59) NL: 0.02 C: 0.23	0.85 (0.73, 0.98) L: 0.03
S3	0.89 (0.75, 1.05) L: 0.17	0.86 (0.75, 0.99) L: 0.03	1.15 (0.89, 1.49) NL: 0.02 C: 0.18	0.85 (0.73, 0.99) L: 0.04
S4	0.93 (0.82, 1.05) L: 0.23	0.91 (0.80, 1.03) L: 0.12	1.27 (1.01, 1.59) NL: 0.003 C: 0.03	1.01 (0.83, 1.23) NL: 0.05 C: 0.95
S5	0.96 (0.89, 1.04) L: 0.36	0.95 (0.88, 1.04) L: 0.25	1.02 (0.94, 1.11) L: 0.69	0.96 (0.87, 1.06) L: 0.38
S6	0.90 (0.76, 1.07) L: 0.23	0.84 (0.74, 0.96) L: 0.02	1.19 (0.83, 1.69) NL: 0.03 C: 0.24	0.81 (0.68, 0.95) L: 0.01

KEY

U=Minimally adjusted analysis
A=Fully adjusted analysis
S1=Missing province-years received $0.04 in annual per capita funding
S2=Missing province-years received $0.23 in annual per capita funding
S3=Missing province-years received $0.32 in annual per capita funding
S4=Missing province-years received $0.56 in annual per capita funding
S5=Missing province-years received $0.93 in annual per capita funding
S6=Inverse probability weighting
■ Funding associated with improved health outcomes
■ Funding significantly associated with improved health outcomes
■ Concave relationship between funding & health outcomes
■ Significant concave relationship between funding & health outcomes
■ Funding associated with worse health outcomes
■ Funding significantly associated with worse health outcomes
☐ Funding not associated with health outcomes

Table 3.7 Summary of sensitivity analyses results for HIV testing at ANC. Cells include the risk ratio (RR) corresponding to a $0.33 increase in annual PEPFAR funding per capita. Table shows p-values for test of linear trend (L) or, if significant departures from linearity were detected, the test for non-linearity (NL) and test for the significance of the curve (C).

	HIV testing at ANC			
	No lag	**1 year lag**	**2 year lag**	**3 year lag**
U	0.99 (0.96, 1.03) L: 0.61	1.01 (0.97, 1.06) L: 0.48	1.01 (0.97, 1.04) L: 0.67	1.00 (0.97, 1.03) L: 0.88
A	0.98 (0.94, 1.01) L: 0.20	1.01 (0.97, 1.05) L: 0.57	1.02 (0.98, 1.06) L: 0.33	1.02 (0.99, 1.05) L: 0.15
S1	0.99 (0.97, 1.02) L: 0.61	0.99 (0.96, 1.02) L: 0.44	0.93 (0.88, 0.99) NL: 0.009 C: 0.02	1.01 (0.99, 1.04) L: 0.26
S2	0.94 (0.88, 1.00) NL: 0.04 C: 0.04	1.00 (0.96, 1.03) L: 0.91	1.02 (0.99, 1.05) L: 0.29	1.02 (0.99, 1.05) L: 0.17
S3	0.99 (0.95, 1.02) L: 0.43	1.01 (0.97, 1.04) L: 0.78	1.03 (1.00, 1.06) L: 0.08	1.02 (0.99, 1.05) L: 0.18
S4	0.99 (0.96, 1.02) L: 0.46	1.02 (0.98, 1.05) L: 0.34	1.04 (1.01, 1.07) L: 0.01	1.01 (0.99, 1.04) L: 0.29
S5	0.99 (0.97, 1.02) L: 0.60	1.01 (0.99, 1.03) L: 0.22	1.03 (1.01, 1.04) L: 0.008	1.01 (0.99, 1.03) L: 0.42

KEY

U=Minimally adjusted analysis
A=Fully adjusted analysis
S1=Missing province-years received $0.04 in annual per capita funding
S2=Missing province-years received $0.23 in annual per capita funding
S3=Missing province-years received $0.32 in annual per capita funding
S4=Missing province-years received $0.56 in annual per capita funding
S5=Missing province-years received $0.93 in annual per capita funding
S6=Inverse probability weighting
■ Funding associated with improved health outcomes
■ Funding significantly associated with improved health outcomes
■ Concave relationship between funding & health outcomes
■ Significant concave relationship between funding & health outcomes
■ Funding associated with worse health outcomes
■ Funding significantly associated with worse health outcomes
☐ Funding not associated with health outcomes

Table 3.8 Summary of sensitivity analyses for neonatal mortality. Cells include the risk ratio (RR) corresponding to a $0.33 increase in annual PEPFAR funding per capita. Table show p-values for test of linear trend (L). No significant departures from linearity were detected.

	Neonatal Mortality			
	No lag	**1 year lag**	**2 year lag**	**3 year lag**
U	1.17 (0.92, 1.49) L: 0.22	1.03 (0.83, 1.28) L: 0.80	0.99 (0.82, 1.20) L: 0.94	0.94 (0.75, 1.18) L: 0.60
A	1.21 (0.96, 1.52) L: 0.12	1.05 (0.84, 1.31) L: 0.66	1.00 (0.83, 1.21) L: 0.98	0.96 (0.77, 1.19) L: 0.69
S1	1.12 (0.92, 1.36) L: 0.28	0.98 (0.81, 1.19) L: 0.86	0.89 (0.75, 1.04) L: 0.14	0.94 (0.80, 1.12) L: 0.50
S2	1.13 (0.91, 1.41) L: 0.28	1.02 (0.84, 1.25) L: 0.82	1.62 (1.07, 2.46) NL: 0.01 C: 0.02	1.01 (0.83, 1.23) L: 0.91
S3	1.11 (0.90, 1.38) L: 0.34	1.04 (0.86, 1.26) L: 0.66	1.57 (1.11, 2.24) NL: 0.008 C: 0.01	1.05 (0.85, 1.29) L: 0.66
S4	1.04 (0.88, 1.23) L: 0.64	1.06 (0.91, 1.22) L: 0.47	1.53 (1.12, 2.08) NL: 0.03 C: 0.01	1.09 (0.90, 1.31) L: 0.38
S5	1.00 (0.90, 1.12) L: 0.94	1.04 (0.95, 1.14) L: 0.43	1.15 (1.03, 1.28) L: 0.02	1.07 (0.95, 1.20) L: 0.29

KEY

U=Minimally adjusted analysis
A=Fully adjusted analysis
S1=Missing province-years received $0.04 in annual per capita funding
S2=Missing province-years received $0.23 in annual per capita funding
S3=Missing province-years received $0.32 in annual per capita funding
S4=Missing province-years received $0.56 in annual per capita funding
S5=Missing province-years received $0.93 in annual per capita funding
S6=Inverse probability weighting
■ Funding associated with improved health outcomes
■ Funding significantly associated with improved health outcomes
■ Concave relationship between funding & health outcomes
■ Significant concave relationship between funding & health outcomes
■ Funding associated with worse health outcomes
■ Funding significantly associated with worse health outcomes
☐ Funding not associated with health outcomes

Table 3.9 Relationship between PEPFAR funding for PMTCT and infant mortality by maternal HIV status among infants among infants with known maternal HIV status based on HIV testing conducted in the 2003 and 2008 surveys. PEPFAR funding was dichotomized as pre- vs. post-PEPFAR.

	Maternal HIV Status		Test for interaction between PEPFAR funding and HIV status
	HIV negative n/N=214/3803 RR, (95% CI)	HIV positive n/N=45/369 RR, (95% CI)	
No lag	0.32 (0.13, 0.77)	0.63 (0.24, 1.67)	0.09
1 year lag	0.53 (0.33, 0.86)	1.00 (0.43, 2.30)	0.15
2 year lag	1.08 (0.62, 1.90)	1.88 (0.73, 4.85)	0.28
3 year lag	0.77 (0.40, 1.46)	1.47 (0.49, 4.37)	0.33

Bibliography

1. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M, et al. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ. 2008;337:a1655.

2. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions: Medical Research Council guidance. BMJ. 2015;350:h1258.

3. Cori A, Ayles H, Beyers N, Schaap A, Floyd S, Sabapathy K, et al. HPTN 071 (PopART): a cluster-randomized trial of the population impact of an HIV combination prevention intervention including universal testing and treatment: mathematical model. PLoS One. 2014;9(1):e84511.

4. Piper ME, Cook JW, Schlam TR, Jorenby DE, Smith SS, Collins LM, et al. A Randomized Controlled Trial of an Optimized Smoking Treatment Delivered in Primary Care. Annals of Behavioral Medicine. 2018.

5. Lloyd J, Creanor S, Logan S, Green C, Dean SG, Hillsdon M, et al. Effectiveness of the Healthy Lifestyles Programme (HeLP) to prevent obesity in UK primary-school children: a cluster randomised controlled trial. The Lancet Child & Adolescent Health. 2018;2(1):35-45.

6. Collins LM, Nahum-Shani I, Almirall D. Optimization of behavioral dynamic treatment regimens based on the sequential, multiple assignment, randomized trial (SMART). Clin Trials. 2014;11(4):426-34.

7. Collins LM, Dziak JJ, Kugler KC, Trail JB. Factorial experiments: efficient tools for evaluation of intervention components. Am J Prev Med. 2014;47(4):498-504.

8. Dziak JJ, Nahum-Shani I, Collins LM. Multilevel factorial experiments for developing behavioral interventions: power, sample size, and resource considerations. Psychol Methods. 2012;17(2):153-75.

9. Collins LM, Chakraborty B, Murphy SA, Strecher V. Comparison of a phased experimental approach and a single randomized clinical trial for developing multicomponent behavioral interventions. Clin Trials. 2009;6(1):5-15.

10. Nevo D, Lok J, Spiegelman D. Analysis of Learn-As-You-Go (LAGO) Studies: Submitted; 2017.

11. Kingston B, Bacallao M, Smokowski P, Sullivan T, Sutherland K. Constructing "Packages" of Evidence-Based Programs to Prevent Youth Violence: Processes and Illustrative Examples From the CDC's Youth Violence Prevention Centers. J Prim Prev. 2016;37(2):141-63.

12. Pettifor A, Nguyen NL, Celum C, Cowan FM, Go V, Hightow-Weidman L. Tailored combination prevention packages and PrEP for young key populations. J Int AIDS Soc. 2015;18(2 Suppl 1):19434.

13. Grant A, Dreischulte T, Guthrie B. Process evaluation of the data-driven quality improvement in primary care (DQIP) trial: active and less active ingredients of a multi-component complex intervention to reduce high-risk primary care prescribing. Implement Sci. 2017;12(1):4.

14. Winder R, Richards SH, Campbell JL, Richards DA, Dickens C, Gandhi M, et al. Development and refinement of a complex intervention within cardiac rehabilitation services: experiences from the CADENCE feasibility study. Pilot Feasibility Stud. 2017;3:9.

15. Shojania KG. Conventional evaluations of improvement interventions: more trials or just more tribulations? BMJ Qual Saf. 2013;22(11):881-4.

16. Hirschhorn LR, Semrau K, Kodkany B, Churchill R, Kapoor A, Spector J, et al. Learning before leaping: integration of an adaptive study design process prior to initiation of BetterBirth, a large-scale randomized controlled trial in Uttar Pradesh, India. Implement Sci. 2015;10:117.

17. Kara N, Firestone R, Kalita T, Gawande AA, Kumar V, Kodkany B, et al. The BetterBirth Program: Pursuing Effective Adoption and Sustained Use of the WHO Safe Childbirth Checklist Through Coaching-Based Implementation in Uttar Pradesh, India. Global Health: Science and Practice. 2017;5(2):232-43.

18. Spector JM, Lashoher A, Agrawal P, Lemer C, Dziekan G, Bahl R, et al. Designing the WHO Safe Childbirth Checklist program to improve quality of care at childbirth. Int J Gynaecol Obstet. 2013;122(2):164-8.

19. Haynes AB, Berry WR, Gawande AA. What do we know about the safe surgery checklist now? Ann Surg. 2015;261(5):829-30.

20. Haynes AB, Weiser TG, Berry WR, Lipsitz SR, Breizat AH, Dellinger EP, et al. A surgical safety checklist to reduce morbidity and mortality in a global population. N Engl J Med. 2009;360(5):491-9.

21. Spector JM, Agrawal P, Kodkany B, Lipsitz S, Lashoher A, Dziekan G, et al. Improving quality of care for maternal and newborn health: prospective pilot study of the WHO safe childbirth checklist program. PLoS One. 2012;7(5):e35151.

22. Semrau KEA, Hirschhorn LR, Kodkany B, Spector JM, Tuller DE, King G, et al. Effectiveness of the WHO Safe Childbirth Checklist program in reducing severe maternal, fetal, and newborn harm in Uttar Pradesh, India: study protocol for a matched-pair, cluster-randomized controlled trial. Trials. 2016;17(1):576.

23. Fitzmaurice G, Laird N, Ware J. Applied Longitudinal Analysis. 2 ed. Hoboken, New Jersey: John Wiley & Sons, Inc; 2011.

24. Durrleman S, Simon R. Flexible regression models with cubic splines. Stat Med. 1989;8(5):551-61.

25. Weiss CH. Nothing as Practical as Good Theory: Exploring Theory-Based Evaluation for Comprehensive Community Initiatives for Children and Families. In: Connell J, Kubisch A, Schorr L, Weiss C, editors. New approaches to evaluating community initiatives: Concepts, methods, and contexts. Washington DC: Aspen Institute; 1995. p. 65-92.

26. De Silva MJ, Breuer E, Lee L, Asher L, Chowdhary N, Lund C, et al. Theory of Change: a theory-driven approach to enhance the Medical Research Council's framework for complex interventions. Trials. 2014;15:267.

27. Breuer E, Lee L, De Silva M, Lund C. Using theory of change to design and evaluate public health interventions: a systematic review. Implement Sci. 2016;11:63.

28. Liang L, Bernhardsson S, Vernooij RW, Armstrong MJ, Bussieres A, Brouwers MC, et al. Use of theory to plan or evaluate guideline implementation among physicians: a scoping review. Implement Sci. 2017;12(1):26.

29. Ajzen I. Attitudes, personality, and behavior: McGraw-Hill Education (UK); 2005.

30. Atkins L, Francis J, Islam R, O'Connor D, Patey A, Ivers N, et al. A guide to using the Theoretical Domains Framework of behaviour change to investigate implementation problems. Implement Sci. 2017;12(1):77.

31. Band R, Bradbury K, Morton K, May C, Michie S, Mair FS, et al. Intervention planning for a digital intervention for self-management of hypertension: a theory-, evidence- and person-based approach. Implement Sci. 2017;12(1):25.

32. Hallinan CM. Program logic: a framework for health program design and evaluation - the Pap nurse in general practice program. Aust J Prim Health. 2010;16(4):319-25.

33. Tremblay MC, Brousselle A, Richard L, Beaudet N. Defining, illustrating and reflecting on logic analysis with an example from a professional development program. Eval Program Plann. 2013;40:64-73.

34. Mackenzie M, Blamey A. The Practice and the Theory. Evaluation. 2016;11(2):151-68.

35. Chandani Y, Noel M, Pomeroy A, Andersson S, Pahl MK, Williams T. Factors affecting availability of essential medicines among community health workers in Ethiopia, Malawi, and Rwanda: solving the last mile puzzle. Am J Trop Med Hyg. 2012;87(5 Suppl):120-6.

36. de Wit EE, Adithy, Chakranarayan C, Bunders-Aelen JFG, Regeer BJ. Learning About Parenting Together: A Programme to Support Parents with Inter-generational Concerns in Pune, India. Contemp Fam Ther. 2018;40(1):68-83.

37. Victora CG, Black RE, Boerma JT, Bryce J. Measuring impact in the Millennium Development Goal era and beyond: a new approach to large-scale effectiveness evaluations. The Lancet. 2011;377(9759):85-95.

38. Gooding K, Makwinja R, Nyirenda D, Vincent R, Sambakunsi R. Using theories of change to design monitoring and evaluation of community engagement in research: experiences from a research institute in Malawi. Wellcome Open Res. 2018;3:8.

39. Gilissen J, Pivodic L, Gastmans C, Vander Stichele R, Deliens L, Breuer E, et al. How to achieve the desired outcomes of advance care planning in nursing homes: a theory of change. BMC Geriatr. 2018;18(1):47.

40. Pronovost PJ, Berenholtz SM, Needham DM. Translating evidence into practice: a model for large scale knowledge translation. BMJ. 2008;337:a1714.

41. Molina RL, Bobay L, Semrau KEA. Historical Perspectives: Lessons from the BetterBirth Trial: A Practical Roadmap for Complex Intervention Studies. NeoReviews. 2019;20(2):e62-e6.

42. Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios LP, et al. A tutorial on pilot studies: the what, why and how. BMC Med Res Methodol. 2010;10:1.

43. Lee EC, Whitehead AL, Jacques RM, Julious SA. The statistical interpretation of pilot trials: should significance thresholds be reconsidered? BMC Med Res Methodol. 2014;14:41.

44. Prentice RL. Surrogate endpoints in clinical trials: Definition and operational criteria. Statistics in Medicine. 1989;8(4):431-40.

45. Vanderweele TJ. Surrogate measures and consistent surrogates. Biometrics. 2013;69(3):561-9.

46. Chen H, Geng Z, Jia J. Criteria for surrogate end points. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2007;69(5):919-32.

47. Ciani O, Buyse M, Garside R, Pavey T, Stein K, Sterne JA, et al. Comparison of treatment effect sizes associated with surrogate and final patient relevant outcomes in randomised controlled trials: meta-epidemiological study. BMJ. 2013;346:f457.

48. Fleming TR, DeMets DL. Surrogate End Points in Clinical Trials: Are We Being Misled? Annals of Internal Medicine. 1996;125(7):605.

49. Packer M, Pitt B, Rouleau JL, Swedberg K, DeMets DL, Fisher L. Long-Term Effects of Flosequinan on the Morbidity and Mortality of Patients With Severe Chronic Heart Failure: Primary Results of the PROFILE Trial After 24 Years. JACC Heart Fail. 2017;5(6):399-407.

50. Concorde Coordinating Committee. Preliminary analysis of the Concorde trial. Lancet. 1993.

51. Packer M, Carver JR, Rodeheffer RJ, Ivanhoe RJ, DiBianco R, Zeldis SM, et al. Effect of oral milrinone on mortality in severe chronic heart failure. The PROMISE Study Research Group. N Engl J Med. 1991;325(21):1468-75.

52. Gilbert PB, Qin L, Self SG. Evaluating a surrogate endpoint at three levels, with application to vaccine development. Stat Med. 2008;27(23):4758-78.

53. Ensor H, Lee RJ, Sudlow C, Weir CJ. Statistical approaches for evaluating surrogate outcomes in clinical trials: A systematic review. J Biopharm Stat. 2016;26(5):859-79.

54. Chuang-Stein C, Kirby S, Hirsch I, Atkinson G. The role of the minimum clinically important difference and its impact on designing a trial. Pharm Stat. 2011;10(3):250-6.

55. Arnold DM, Burns KE, Adhikari NK, Kho ME, Meade MO, Cook DJ, et al. The design and interpretation of pilot trials in clinical research in critical care. Crit Care Med. 2009;37(1 Suppl):S69-74.

56. Semrau KEA, Hirschhorn LR, Marx Delaney M, Singh VP, Saurastri R, Sharma N, et al. Outcomes of a Coaching-Based WHO Safe Childbirth Checklist Program in India. N Engl J Med. 2017;377(24):2313-24.

57. Semrau KEA, Miller K, Lipsitz S, Fisher-Bowman J, Karlage A, Neville BA, et al., editors. Association of Adherence to Essential Birth Practices and Perinatal Mortality in Uttar Pradesh,

India. Oral Presentation (Number 766). Federation of International Gynecologists & Obstetricians (FIGO) 22nd World Congress; October 18 2018; Rio de Janeiro.

58. Hargreaves JR, Goodman C, Davey C, Willey BA, Avan BI, Schellenberg JR. Measuring implementation strength: lessons from the evaluation of public health strategies in low- and middle-income settings. Health Policy Plan. 2016;31(7):860-7.

59. Warren SF, Fey ME, Yoder PJ. Differential treatment intensity research: a missing link to creating optimally effective communication interventions. Ment Retard Dev Disabil Res Rev. 2007;13(1):70-7.

60. Fletcher A, Jamal F, Moore G, Evans RE, Murphy S, Bonell C. Realist complex intervention science: Applying realist principles across all phases of the Medical Research Council framework for developing and evaluating complex interventions. Evaluation (Lond). 2016;22(3):286-303.

61. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. BMJ. 2015;350:h391.

62. Nelson MC, Cordray DS, Hulleman CS, Darrow CL, Sommer EC. A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. J Behav Health Serv Res. 2012;39(4):374-96.

63. Carroll C, Patterson M, Wood S, Booth A, Rick J, Balain S. A conceptual framework for implementation fidelity. Implement Sci. 2007;2:40.

64. Hasson H. Systematic evaluation of implementation fidelity of complex interventions in health and social care. Implement Sci. 2010;5:67.

65. Breitenstein SM, Gross D, Garvey CA, Hill C, Fogg L, Resnick B. Implementation fidelity in community-based interventions. Res Nurs Health. 2010;33(2):164-73.

66. Greenland S. Randomization, statistics, and causal inference. Epidemiology. 1990;1(6):421-9.

67. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol. 1986;15(3):413-9.

68. Spiegelman D, Zhou X. Evaluating Public Health Interventions: 8. Causal Inference for Time-Invariant Interventions. Am J Public Health. 2018:e1-e4.

69. Piper ME, Fiore MC, Smith SS, Fraser D, Bolt DM, Collins LM, et al. Identifying effective intervention components for smoking cessation: a factorial screening experiment. Addiction. 2016;111(1):129-41.

70. Hermens RP, Hak E, Hulscher ME, Braspenning JC, Grol RP. Adherence to guidelines on cervical cancer screening in general practice: programme elements of successful implementation. Br J Gen Pract. 2001;51(472):897-903.

71. Pellecchia M, Connell JE, Beidas RS, Xie M, Marcus SC, Mandell DS. Dismantling the Active Ingredients of an Intervention for Children with Autism. J Autism Dev Disord. 2015;45(9):2917-27.

72. Abry T, Hulleman CS, Rimm-Kaufman SE. Using Indices of Fidelity to Intervention Core Components to Identify Program Active Ingredients. American Journal of Evaluation. 2015;36(3):320-38.

73. Prins A, Oulmette P, Kimerling R, Cameron RP, Hugelshofer DS, Shaw-Hegwer J, et al. The primary care PTSD screen (PD-PTSD): development and operating characteristics. Primary Care Psychiatry. 2003;9(1):9-14.

74. Werdenberg J, Biziyaremye F, Nyishime M, Nahimana E, Mutaganzwa C, Tugizimana D, et al. Successful implementation of a combined learning collaborative and mentoring intervention to improve neonatal quality of care in rural Rwanda. BMC Health Serv Res. 2018;18(1):941.

75. Semrau KEA, Herlihy J, Grogan C, Musokotwane K, Yeboah-Antwi K, Mbewe R, et al. Effectiveness of 4% chlorhexidine umbilical cord care on neonatal mortality in Southern Province, Zambia (ZamCAT): a cluster-randomised controlled trial. The Lancet Global Health. 2016;4(11):e827-e36.

76. Wyatt KM, Lloyd JJ, Creanor S, Logan S. The development, feasibility and acceptability of a school-based obesity prevention programme: results from three phases of piloting. BMJ Open. 2011;1(1):e000026.

77. Richards SH, Dickens C, Anderson R, Richards DA, Taylor RS, Ukoumunne OC, et al. Assessing the effectiveness of Enhanced Psychological Care for patients with depressive symptoms attending cardiac rehabilitation compared with treatment as usual (CADENCE): a pilot cluster randomised controlled trial. Trials. 2018;19(1):211.

78. Alkema L, Chou D, Hogan D, Zhang S, Moller A-B, Gemmill A, et al. Global, regional, and national levels and trends in maternal mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Maternal Mortality Estimation Inter-agency Group. The Lancet. 2016;387(10017):462-74.

79. Wang H, Bhutta ZA, Coates MM, Coggeshall M, Dandona L, Diallo K, et al. Global, regional, national, and selected subnational levels of stillbirths, neonatal, infant, and under-5 mortality, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. The Lancet. 2016;388(10053):1725-74.

80. Ahmed I, Ali SM, Amenga-Etego S, Ariff S, Bahl R, Baqui AH, et al. Population-based rates, timing, and causes of maternal deaths, stillbirths, and neonatal deaths in south Asia and sub-Saharan Africa: a multi-country prospective cohort study. The Lancet Global Health. 2018;6(12):e1297-e308.

81. Campbell OMR, Calvert C, Testa A, Strehlow M, Benova L, Keyes E, et al. The scale, scope, coverage, and capability of childbirth care. The Lancet. 2016;388(10056):2193-208.

82. Harvey S. Are skilled birth attendants really skilled? A measurement method, some disturbing results and a potential way forward. Bulletin of the World Health Organization. 2007;85(10):783-90.

83. Randive B, Diwan V, De Costa A. India's Conditional Cash Transfer Programme (the JSY) to Promote Institutional Birth: Is There an Association between Institutional Birth Proportion and Maternal Mortality? PLoS One. 2013;8(6):e67452.

84. Nagpal J, Sachdeva A, Sengupta Dhar R, Bhargava VL, Bhartia A. Widespread non-adherence to evidence-based maternity care guidelines: a population-based cluster randomised household survey. BJOG: An International Journal of Obstetrics & Gynaecology. 2015;122(2):238-47.

85. Bhutta ZA, Das JK, Bahl R, Lawn JE, Salam RA, Paul VK, et al. Can available interventions end preventable deaths in mothers, newborn babies, and stillbirths, and at what cost? The Lancet. 2014;384(9940):347-70.

86. Pariyo GW, Gouws E, Bryce J, Burnham G, Uganda IMCI Impact Study Team. Improving facility-based care for sick children in Uganda: training is not enough. Health Policy Plan. 2005;20 Suppl 1:i58-i68.

87. Ersdal HL, Vossius C, Bayo E, Mduma E, Perlman J, Lippert A, et al. A one-day "Helping Babies Breathe" course improves simulated performance but not clinical management of neonates. Resuscitation. 2013;84(10):1422-7.

88. Schwellnus H, Carnahan H. Peer-coaching with health care professionals: what is the current status of the literature and what are the key components necessary in peer-coaching? A scoping review. Med Teach. 2014;36(1):38-46.

89. Thompson R, Wolf DM, Sabatine JM. Mentoring and coaching: a model guiding professional nurses to executive success. J Nurs Adm. 2012;42(11):536-41.

90. Prasad S, Sopdie E, Meya D, Kalbarczyk A, Garcia PJ. Conceptual framework of mentoring in low- and middle-income countries to advance global health. Am J Trop Med Hyg. 2019;100(1_Suppl):9-14.

91. Pearson M, Brew A. Research Training and Supervision Development. Studies in Higher Education. 2002;27(2):135-50.

92. Rowe AK, de Savigny D, Lanata CF, Victora CG. How can we achieve and maintain high-quality performance of health workers in low-resource settings? The Lancet. 2005;366(9490):1026-35.

93. Rowe AK, Rowe SY, Peters DH, Holloway KA, Chalker J, Ross-Degnan D. Effectiveness of strategies to improve health-care provider practices in low-income and middle-income countries: a systematic review. The Lancet Global Health. 2018;6(11):e1163-e75.

94. Magge H, Anatole M, Cyamatare FR, Mezzacappa C, Nkikabahizi F, Niyonzima S, et al. Mentoring and quality improvement strengthen integrated management of childhood illness implementation in rural Rwanda. Arch Dis Child. 2015;100(6):565-70.

95. Rowe AK, Onikpo F, Lama M, Osterholt DM, Rowe SY, Deming MS. A multifaceted intervention to improve health worker adherence to integrated management of childhood illness guidelines in Benin. Am J Public Health. 2009;99(5):837-46.

96. Trap B, Todd CH, Moore H, Laing R. The impact of supervision on stock management and adherence to treatment guidelines: a randomized controlled trial. Health Policy Plan. 2001;16(3):273-80.

97. Chalker J. Improving antibiotic prescribing in Hai Phong Province, Viet Nam: the ‘‘antibiotic-dose’’ indicator. Bulletin of the World Health Organization. 2001;79(4):313-20.

98. Loevinsohn BP, Guerrero ET, Gregorio SP. Improving primary health care through systematic supervision: a controlled field trial. Health Policy and Planning. 1995;10(2):144-53.

99. Bello DA, Hassan ZI, Afolaranmi TO, Tagurum YO, Chirdan OO, Zoakah AI. Supportive supervision: An effective intervention in achieving high quality malaria case management at primary health care level in Jos, Nigeria. Annals of African Medicine. 2013;12(4):243-51.

100. Broughton EI, Karamagi E, Kigonya A, Lawino A, Marquez L, Lunsford SS, et al. The cost-effectiveness of three methods of disseminating information to improve medical male circumcision in Uganda. PLoS One. 2018;13(4):e0195691.

101. Stanback J, Griffey S, Lynam P, Ruto C, Cummings S. Improving adherence to family planning guidelines in Kenya: an experiment. Int J Qual Health Care. 2007;19(2):68-73.

102. Manzi A, Nyirazinyoye L, Ntaganira J, Magge H, Bigirimana E, Mukanzabikeshimana L, et al. Beyond coverage: improving the quality of antenatal care delivery through integrated mentorship and quality improvement at health centers in rural Rwanda. BMC Health Serv Res. 2018;18(1):136.

103. Hirschhorn LR, Krasne M, Maisonneuve J, Kara N, Kalita T, Henrich N, et al. Integration of the Opportunity-Ability-Motivation behavior change framework into a coaching-based WHO Safe Childbirth Checklist program in India. Int J Gynaecol Obstet. 2018;142(3):321-8.

104. Govindaraju R, Hadining A, Chandra D. Physicians' Adoption of Electronic Medical Records: Model Development Using Ability – Motivation - Opportunity Framework. In: Mustofa K, Neuhold EJ, Tjoa AM, Weippl E, You I, editors. Information and Communication Technology: International Conference, ICT-EurAsia 2013. Lecture Notes in Computer Science. Yogyakarta, Indonesia: Springer; 2013.

105. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci. 2011;6:42.

106. Annual Health Survey 2012-2013 Fact Sheet: Uttar Pradesh. India: Office of the Registrar General & Census Commissioner.

107. Gass JD, Jr., Misra A, Yadav MNS, Sana F, Singh C, Mankar A, et al. Implementation and results of an integrated data quality assurance protocol in a randomized controlled trial in Uttar Pradesh, India. Trials. 2017;18(1):418.

108. Wacholder S. Binomial regression in GLIM: estimating risk ratios and risk differences. Am J Epidemiol. 1986;123(1):174-84.

109. Rotnitzky A, Jewell NP. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. Biometrika. 1990;77(3):485-97.

110. Ory MG, Lee Smith M, Mier N, Wernicke MM. The science of sustaining health behavior change: the health maintenance consortium. Am J Health Behav. 2010;34(6):647-59.

111. Hertzmark E, Li R, Hong B, Spiegelman D. The SAS GLMCURV9 Macro 2014. Available from: https://www.hsph.harvard.edu/donna-spiegelman/software/glmcurv9/.

112. Gomez PP, Nelson AR, Asiedu A, Addo E, Agbodza D, Allen C, et al. Accelerating newborn survival in Ghana through a low-dose, high-frequency health worker training approach: a cluster randomized trial. BMC Pregnancy Childbirth. 2018;18(1):72.

113. Mduma E, Ersdal H, Svensen E, Kidanto H, Auestad B, Perlman J. Frequent brief on-site simulation training and reduction in 24-h neonatal mortality--an educational intervention study. Resuscitation. 2015;93:1-7.

114. Patabendige M, Senanayake H. Implementation of the WHO safe childbirth checklist program at a tertiary care setting in Sri Lanka: a developing country experience. BMC Pregnancy Childbirth. 2015;15:12.

115. Kabongo L, Gass J, Kivondo B, Kara N, Semrau K, Hirschhorn LR. Implementing the WHO Safe Childbirth Checklist: lessons learnt on a quality improvement initiative to improve mother and newborn care at Gobabis District Hospital, Namibia. BMJ Open Qual. 2017;6(2):e000145.

116. Kumar S, Yadav V, Balasubramaniam S, Jain Y, Joshi CS, Saran K, et al. Effectiveness of the WHO SCC on improving adherence to essential practices during childbirth, in resource constrained settings. BMC Pregnancy Childbirth. 2016;16(1):345.

117. Senanayake HM, Patabendige M, Ramachandran R. Experience with a context-specific modified WHO safe childbirth checklist at two tertiary care settings in Sri Lanka. BMC Pregnancy Childbirth. 2018;18(1):411.

118. Tuyishime E, Park PH, Rouleau D, Livingston P, Banguti PR, Wong R. Implementing the World Health Organization safe childbirth checklist in a district Hospital in Rwanda: a pre- and post-intervention study. Matern Health Neonatol Perinatol. 2018;4:7.

119. Kim B, Miller C, Ritchie M, Smith J, Kirchner J. Time-motion analysis of implementing the collaborative chronic care model in general mental health clinics: Assessing external facilitation effort over time using continuous and interval-based data collection approaches. 11th Annual Conference on the Science of Dissemination and Implementation in Health; December 3-5 2018; Washington, D.C.

120. Mainz J. Developing evidence-based clinical indicators: a state of the art methods primer. Int J Qual Health Care. 2003;15 Suppl 1:i5-11.

121. Hill K, Cheluget B, Curtix S, Bicego G, Mahy M. HIV and increases in childhood mortality in Kenya in the late 1980s to the mid-1990s. United States Agency for International Development (USAID), Measure Evaluation; 2004.

122. Wafula SW, Ikamari LD, K'Oyugi BO. In search for an explanation to the upsurge in infant mortality in Kenya during the 1988-2003 period. BMC Public Health. 2012;12:441.

123. Kenya National AIDS and STI Control Programme (NASCOP). Kenya AIDS Indicator Survey 2012: Final Report. Nairobi: NASCOP; 2014.

124. PEPFAR Dashboards: Office of U.S. Global AIDS Coordinator,. Available from: https://data.pepfar.net/global.

125. Kenya National Bureau of Statistics, Kenya Ministry of Health, Kenya National AIDS Control Council, Kenya Medical Research Institute, Kenya National Council for Population

Development. Kenya Demographic and Health Survey 2014: Final Report. Rockville, MD, USA2015.

126. De Cock KM, Fowler MG, Mercier E, de Vincenzi I, Saba J, Hoff E, et al. Prevention of mother-to-child HIV transmission in resource-poor countries: translating research into policy and practice. JAMA. 2000;283(9):1175-82.

127. Newell ML, Coovadia H, Cortina-Borja M, Rollins N, Gaillard P, Dabis F, et al. Mortality of infected and uninfected infants born to HIV-infected mothers in Africa: a pooled analysis. Lancet. 2004;364(9441):1236-43.

128. Liu L, Oza S, Hogan D, Perin J, Rudan I, Lawn JE, et al. Global, regional, and national causes of child mortality in 2000-13, with projections to inform post-2015 priorities: an updated systematic analysis. Lancet. 2015;385(9966):430-40.

129. Luboga SA, Stover B, Lim TW, Makumbi F, Kiwanuka N, Lubega F, et al. Did PEPFAR investments result in health system strengthening? A retrospective longitudinal study measuring non-HIV health service utilization at the district level. Health Policy Plan. 2016;31(7):897-909.

130. Lee MM, Izama MP. Aid Externalities: Evidence from PEPFAR in Africa. World Development. 2015;67:281-94.

131. Bendavid E, Bhattacharya J. The President's Emergency Plan for AIDS Relief in Africa: An Evaluation of Outcomes. Annals of Internal Medicine. 2009;150(10):688-95.

132. Bendavid E, Holmes CB, Bhattacharya J, Miller G. HIV development assistance and adult mortality in Africa. JAMA. 2012;307(19):2060-7.

133. Cohen RL, Li Y, Giese R, Mancuso JD. An evaluation of the President's Emergency Plan for AIDS Relief effect on health systems strengthening in sub-Saharan Africa. J Acquir Immune Defic Syndr. 2013;62(4):471-9.

134. Duber HC, Coates TJ, Szekeras G, Kaji AH, Lewis RJ. Is there an association between PEPFAR funding and improvement in national health indicators in Africa? A retrospective study. J Int AIDS Soc. 2010;13:21.

135. Central Bureau of Statistics - CBS/Kenya, Ministry of Health - MOH/Kenya, ORC Macro. Kenya Demographic and Health Survey 2003. Calverton, Maryland, USA: CBS, MOH, and ORC Macro; 2004.

136. Kenya National AIDS and STI Control Programme (NASCOP). Kenya AIDS Indicator Survey 2007: Final Report. Nairobi: NASCOP; 2009.

137. Kenya National Bureau of Statistics, Ministry of Health/Kenya, National AIDS Control Council/Kenya, Kenya Medical Research Institute, Population NCf, Development/Kenya. Kenya Demographic and Health Survey 2014. Rockville, MD, USA; 2015.

138. Kenya National Bureau of Statistics - KNBS, National AIDS Control Council/Kenya, National AIDS/STD Control Programme/Kenya, Health MoP, Sanitation/Kenya, Kenya Medical Research Institute. Kenya Demographic and Health Survey 2008-09. Calverton, Maryland, USA: KNBS and ICF Macro; 2010.

139. Hallett TB, Gregson S, Kurwa F, Garnett GP, Dube S, Chawira G, et al. Measuring and correcting biased child mortality statistics in countries with generalized epidemics of HIV infection. Bull World Health Organ. 2010;88(10):761-8.

140. Zaba B, Marston M, Floyd S. The effect of HIV on Child Mortality Trends in Sub-Saharan Africa. United Nations Population Division, Department of Economic and Social Affairs; 2003.

141. Country Operational Plans: Office of the U.S. Global AIDS Coordinator. Available from: https://www.pepfar.gov/countries/cop/index.htm.

142. Habicht JP, Victora CG, Vaughan JP. Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact. Int J Epidemiol. 1999;28(1):10-8.

143. Subnational Population Database: World Bank Group. Available from: https://data.worldbank.org/data-catalog/subnational-population.

144. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res. 2013;22(3):278-95.

145. Wagner Z, Barofsky J, Sood N. PEPFAR Funding Associated With An Increase In Employment Among Males in Ten Sub-Saharan African Countries. Health Aff (Millwood). 2015;34(6):946-53.

146. Chin RJ, Sangmanee D, Piergallini L. PEPFAR funding and reduction in HIV infection rates in 12 focus sub-Saharan African countries: A quantitative analysis. International Journal of MCH and AIDS. 2015;3(2):150-8.

147. Lima VD, Granich R, Phillips P, Williams B, Montaner JS. Potential impact of the US President's Emergency Plan for AIDS relief on the tuberculosis/HIV coepidemic in selected Sub-Saharan African countries. J Infect Dis. 2013;208(12):2075-84.

148. U.S. State Department, Office of the U.S Global AIDS Coordinator and Health Diplomacy. PEPFAR: 2017 Annual Report to Congress. 2017.

149. Lo NC, Lowe A, Bendavid E. Abstinence funding was not associated with reductions in HIV risk behavior in sub-Saharan Africa. Health Aff (Millwood). 2016;35(5):856-63.

150. ICF International. Demographic and Health Survey Sampling and Household Listing Manual. Calverton, Maryland, U.S.A.: ICF International; 2012.

151. Zou G. A modified poisson regression approach to prospective studies with binary data. Am J Epidemiol. 2004;159(7):702-6.

152. Staveteig S, Mallick L. Intertemporal comparisons of poverty and wealth with DHS data: A harmonized asset index approach. Rockville, Maryland, USA: ICF International; 2014.