Results-Based Financing for Health Impact Evaluation in Burkina Faso

Results report

Version 2 (June 22nd, 2018)

Authors (alphabetical):

Manuela De Allegri, Julia Lohmann, and Michael Schleicher, with contributions from Stephan Brenner and Jean-Louis Koulidiati

Institute of Public Health, Heidelberg University

Contents

Abbreviations

ANC	Antenatal care consultation
CBHI	Community-based health insurance
CHR	Centre Hospitalier Régional (regional referral hospitals)
CMA	Centre Médical avec Antenne chirurgicale (district hospital)
CSPS	Centre de Santé et de Promotion Sociale (health center)
DHMT	District health management team
DID	Difference-in-Differences
IMCI	Integrated management of childhood illness
IUD	Intra-uterine device
MNCH	Maternal, newborn, and child health
MoH	Ministry of Health
PBF	Performance-based financing
PMTCT	Prevention of mother to child transmission (of HIV)
PNC	Postnatal care consultation
SNIS	Système national d'information sanitaire (health information system)
U5	Children under the age of 5 years

0. Executive summary

This report presents key results of the Results-Based Financing for Health Impact Evaluation in Burkina Faso. The impact analyses and the writing of this report were commissioned as an independent evaluation to a team from the Health Economics and Financing Group of the Institute of Public Health, Heidelberg University (HIPH). Design decisions were made jointly with World Bank staff and data collection for the impact evaluation was carried out by Centre MURAZ in Burkina Faso.

In spite of substantial improvements over the course of the last years, Burkina Faso still largely lags behind regional averages on many health indicators, particularly such related to Maternal, Newborn and Child Health (MNCH). Further, inequalities and low accessibility of services still exist, particularly for lower income groups for which continued user charges for a variety of essential health services constitute important barriers to access to health care and expose the population to high risk of catastrophic health expenditure.

Performance-based financing (PBF) interventions, which link health care provider payment mechanisms to predefined outputs, have been perceived by many actors as a potential instrument to improve access to health care and quality of care. The World Bank-funded PBF intervention evaluated in this report is part of the HRITF program implemented in many countries worldwide. After a pre-pilot in three districts, the pilot phase extended coverage of existing PBF interventions to twelve additional districts in six regions, with the aim of improving utilization and quality of MNCH. In light of the above-described socio-economic inequities in access to care, the intervention included demand side interventions in addition to the standard supply-side PBF: targeting of the ultra-poor ('indigents') and Community-Based Health Insurance (CBHI).

The objective of this study was to evaluate the impact of the PBF intervention on quality of care and health care utilization for MNCH services, with a focus on equity outcomes by testing whether PBF in combination with demand-side interventions may additionally improve pro-poor access to health care services. The impact evaluation further looked at PBF impact on selected human resources factors as well as key health outcomes (malnutrition, anemia, malaria).

The study was designed as a quasi-experimental study with a nested experimental component. Twelve intervention districts in six regions were purposely selected and two additional control districts identified in the same or neighboring regions. In 8 of the 12 intervention districts, health facilities were randomized into three different intervention modalities: T1, the standard PBF without demand-side component; T2, the standard PBF in T1 plus a systematic targeting of indigents, who were then exempted from user fees, with facilities being reimbursed this loss in user fees through PBF; and T3, the standard PBF plus targeting of indigents as in T2 plus an additional financial incentive to providers to provide care to the targeted indigents. In 2 of the 12 intervention districts, two different intervention modalities were randomized: the

standard PBF T1, and T4, which combines the standard PBF in T1 with the offer of CBHI in which enrolment fees were covered for targeted indigents.

Data were collected from all health facilities in intervention districts as well as in a random sample of health facilities in control districts, using multiple tools including a facility assessment, a structured health worker survey, direct observations of clinical consultations, and patient exit interviews. In the catchment area of each health facility, one village was randomly sample. Within each village, 15 households with at least one currently pregnant woman or woman who had ended a pregnancy in the 24 months prior to the survey were randomly sampled. A comprehensive household survey covered household and household member characteristics and health-related information as well as detailed information on women's pregnancy and birth history and utilization of reproductive health services, on children's utilization of preventive health services, as well as biomarkers (malaria, anemia) and anthropometric measurements for all women of reproductive age and children under the age of 5. Baseline data collection took place between October 2013 and March 2014, endline data collection between April and June 2017. The data available for the impact evaluation were further complemented with routine health system data derived from the Système National d'Information Sanitaire (SNIS). Data was analyzed using a Difference-in-Differences approach.

The following table summarizes key findings.

	PBF impact in relation to status quo	**Additional value of demand-side components?**
Human resources factors	- No impact on individual performance evaluation - Little impact on health worker satisfaction and perceived agency - No crowding out of intrinsic motivation apparent	- No impact detected
Health service quality	- Positive impact on availability of key infrastructure and ANC routine services - No effects on drug availability - Negative impact on quality of child care services (driven by improvements in controls) - Negative impact on perceived quality of care, particularly ANC (driven by improvements in controls)	- Limited impact detected
Health service utilization	- Positive impact on utilization of maternal care services, par. delivery and PNC services, and family planning among the poorest 20% - Some positive impact on child and adult consultation - No effect on vaccination of children - Mixed effects on growth monitoring	- Limited overall - T2 and T3 less effective than T1 for selected maternal care and preventive child services - T4 more effective than T1 for curative consultations
Population health status	- No impact except for reduction of severe acute child malnutrition among the poorest 20%	- Limited impact detected

The results must be interpreted in light of the implementation context, implementation challenges, and methodological challenges. In regards to implementation context, most importantly, a nation-wide user fee exemption policy (gratuité) was introduced in June 2016, covering many of services incentivized by PBF. While we do expect the implementation of the gratuité not to have interfered with the identification of the effects attributable to PBF (given its national roll out across intervention and control districts), it would be naïf to assume lack of interaction between PBF and such a major health financing reform. For instance, it is possible that healthcare providers focused on provision of services for which both the gratuité and the PBF program provided an explicit financial incentive (e.g. maternal care and curative care to children and pregnant and lactating women), whereas services which have long been free to patients at point of service might have received less attention (e.g. preventive child services). Further, it is possible that PBF could bear a greater impact on services that had long been the target of national policies due to a certain readiness among healthcare providers to enable change. Similarly, it is not surprising that we observed hardly any additional benefit of the intervention arms combining PBF with equity measures. These equity measures were in fact removed for all services included in the gratuité, hence effectively equating the more complex PBF arms to the standard PBF in T1. Further qualitative research is necessary to gain a better understanding of such possible interactions.

In regards to implementation, narrative evidence from implementers as well as the results of a parallel process evaluation led by the University of Montreal underline various implementation challenges which have likely hampered intervention effects. Such challenges include for instance substantial delays in payment resulting in frustration among providers, delays and budgetary limitations in regards to the contract management and verification agents, and unintended dynamics introduced by the indigent selection process and community verification.

In regards to methodology, the study is limited by low statistical power to detect impact of PBF in relation to status quo, due to a relatively small number of districts (i.e. treatment allocation unit). Further, a fundamental prerequisite of inference of intervention impact is that no treatment similar to the intervention in question should have taken place at the same time. In a context in which a multitude of donors and non-governmental organizations are active in addition to government operations, this was impossible to achieve. Further research is necessary to gauge the extent to which effect estimates might in part reflect such other on-going interventions.

1. Background

1.1. Country context

Burkina Faso is a landlocked country located in West Africa, with a population of about 18.5 million. At the time the PBF was first planned, with a purchasing-power-parity (PPP) GDP of USD 1560 in 2013, the country was among the poorest in the world. The 2014 UNDP Human Development Index ranked Burkina Faso 181 out of 186 countries, suggesting no substantial improvement being made in recent years. Nearly half of the population lives below the poverty level, on less than USD 1.25 per day.

In spite of substantial improvements over the course of the last few years, health indicators still largely lag behind regional averages. Life expectancy is at 58 years. In 2013, maternal and under-five mortality were estimated at 400/100,000 and 98/1000 respectively. Malaria, acute respiratory infections, and diarrhea still account for the largest proportion of child mortality, often coupled with an underlying situation of malnutrition, with nearly 40% of all children being classified as stunted.

The health sector suffers from this generalized lack of resources. In 2013, total per capita health expenditure was estimated at 6.4% of GDP, equivalent to PPP USD 109. Government expenditure amounted to 58% of total health expenditure, including contributions by development partners being estimated at 23% of this total. More than 80% of all private expenditure on health is not channeled through pre-paid and pooled mechanisms, exposing the population, especially the poor living in rural areas, to a high risk of catastrophic health expenditure [1][2]. User charges continue to be applied across a variety of essential healthcare services.

Health service delivery is organized according to a three-tier system, with primary facilities (Centre de Santé et Promotion Sociale - CSPS) located in rural areas; district hospitals located in each district capital; and regional and national referral hospital located in the region capitals and in the national capital Ouagadougou. Most health service provision is ensured directly by public facilities, with private provision remaining a relatively small reality, confined to for-profit providers in the main urban centers and religious not-for-profit providers in some rural areas.

The literature has consistently reported that geographical and financial barriers, due to scarcity of facilities in most rural areas and to the imposition of user fees, continue to hamper access to healthcare services [3][4][5][6][7]. The poor health outcomes described above are largely the result of poor access to services, with people largely under-utilizing the care they need.

1.2. History of Performance-based Financing in Burkina Faso

The PBF program at the core of our impact evaluation rests on the experience and knowledge acquired during the implementation of a pre-pilot PBF intervention, managed by the Ministry of Health (MoH) with financial and technical assistance from

the World Bank in the period 2011-2013 in three districts (Titao, Leo, and Boulsa). Within the framework of this early PBF intervention, health facilities and the MoH entered a contractual agreement whereby the MoH would reward the provision of a defined benefit package according to a case-based payment modality, adjusted for quality of service provision. Each quarter, prior to payments being made, an external agency was engaged to verify both the quantity and the quality of the services provided. Facilities were granted full autonomy over the newly acquired funds. No explicit provision mandated health providers to devote a specific portion of the additional revenues towards facility upgrades.

An independent evaluation detected a positive effect of the intervention across maternal healthcare services [8] and a marked improvement in the quality of services provided according to Integrated Management of Childhood Illnesses (IMCI) guidelines. Nevertheless, the early PBF intervention proved not to be effective in reaching the most vulnerable sectors of society, resulting in the perpetuation of existing inequities in access to health services across people of different socio-economic status.

In the light of the positive results produced by the early pre-pilot, but acknowledging the difficulties intrinsically related to reaching the very poor, the MoH, again with financial and technical assistance from the World Bank, decided to scale up the PBF intervention to an additional 12 districts, but to do so by combining the standard supply-side intervention with a series of equity measures aimed at overcoming inequalities in access and service provision across socio-economic strata. To understand this decision, it is important to locate the PBF program in Burkina Faso within the broader context of PBF programs supported by the World Bank over the course of the past decade across sub-Saharan countries. By the time plans to expand PBF beyond the initial pre-pilot were undertaken, evidence was emerging from other settings on the potential of PBF to stimulate changes in service provision, but also on its inability to stimulate equity changes if implemented exclusively as a supply-side intervention.

It is also in the light of these considerations that the MoH and its development partners opted to implement PBF in conjunction with a series of equity measures aimed at maximizing the potential of PBF to act as a catalyst for equity changes. It is at this point, building on the knowledge generated in other settings and looking at the specific need to address equity gaps in country, that the World Bank realized the potential to use the case of Burkina Faso to test novel PBF models, combining elements of supply and demand side interventions into a single program. It is also at this point that knowledge generation was first conceptualized as an intrinsic component of the PBF program implementation and that the decision to contract an independent academic institution to carry out the impact evaluation was made. It ought to be noted explicitly that the research team did not have the ability to influence the intervention design, but was involved at a sufficiently early stage to influence roll out and evaluation decisions.

1.3. Intervention design

In line with the objectives of the pre-pilot, the primary objective of the PBF program at the core of our evaluation was to improve the utilization and quality of MNCH services, in particular among vulnerable populations, such as the ultra-poor. Effectively, however, the PBF benefit package was very comprehensive, comprising a broad range of primary and secondary services beyond MNCH, including also general adult curative consultations, HIV and tuberculosis services. **Table 1** contains the list of indicators for primary-level facilities. In line with the considerations outlined above, the MoH decided to implement PBF according to four different models, three of which included special provisions to improve access to care for the ultra-poor.

Table 1: Quantitative indicators for primary-level health care facilities[1]

1	Number of new patients age 5 or older in curative consultation
2	Number of new patients under the age of 5 in curative consultation
3	Number of days of hospitalization
4	Number of counter-references received
5*	Number of children fully vaccinated
6*	Number of pregnant women who have received two or more doses of tetanus vaccine
7*	Number of pregnant women (new and repeat visits) in antenatal care consultation
8*	Number of women in postnatal consultation (6-8 days and 6-8 weeks post-delivery)
9	Number of deliveries performed
10	Number of women (new and repeat visits) in family planning consultation using oral or injectable contraceptives
11	Number of women (new and repeat visits) in family planning consultation using long-term methods (IUD or implant)
12*	Number of new patients aged 0-11 months in growth monitoring consultation
13*	Number of patients aged 12-23 months in growth monitoring consultation
14*	Number of children aged 6-59 months treated for moderate acute malnutrition
15*	Number of children aged 6-59 months treated for severe acute malnutrition without complications
16	*Number of home visits effected (not implemented)*
17*	Number of clients having benefitted from voluntary HIV testing and counselling (excluding pregnant women tested in the context of PMTCT)
18*	Number of pregnant women having benefitted from voluntary HIV testing and counselling in the context of PMTCT
19*	Number of HIV-positive mothers having benefitted from complete prophylactic anti-retroviral treatment
20*	Number of newborns to HIV-positive mothers treated
21*	Number of people living with HIV under anti-retroviral treatment
22	Number of pulmonary tuberculosis cases (new and relapse) detected
23	Number of tuberculosis cases (all types) treated and declared cured or treatment terminated

[1] This initial indicator list evolved slightly in the course of the implementation period under evaluation; indicators marked with a star were not reimbursed at a higher price for indigents as they were covered by the gratuité

To understand the rationale of the intervention, it ought to be specified that the PBF program was conceived and implemented at a time when service delivery in Burkina Faso was still heavily reliant on direct user charges at point of use for the vast majority of primary and secondary services. In June 2016, an exemption policy targeting pregnant and lactating women and children under five came into effect, requiring some adjustment to the unit prices the PBF program applied, but all other services remained available only upon payment at point of use.

Hereafter, we describe the four PBF models in detail:

i. *T1*: Traditional PBF. PBF contracts were signed between the MoH and the health facilities (primary- and secondary-level), indicating a list of services purchased by PBF ('quantity indicators', Table 1 and [9]). External reviewers assess facility report on quantity indicators on a monthly basis. Based on these verified results, contracted facilities received case-based payments in partial reimbursement for the services delivered. Unit prices were calculated a prior by the implementation team, on the basis of the relative cost and frequency of the services provided. Payments were further adjusted for quality of service delivery. Quality was assessed with comprehensive quality checklists [9], verified on a quarterly basis by District Health Management Teams (DHMTs). Facilities received an additional quality bonus calculated on the basis of the quantity outcomes if they achieved a quality score above 50%. T1 did not include any specific provision to facilitate access to care for the ultra-poor.

ii. *T2*: Traditional PBF + systematic targeting and health service subsidization for the ultra-poor. T2 operates PBF according to the same contractual model of T1, but combines it with specific provisions to facilitate access to care for extremely vulnerable households residing in the health facility catchment area. These provisions include:

a. A systematic targeting of the ultra-poor implemented using a community targeting approach. The aim was to identify up to 20% of all households residing in the health facility catchment area as extremely vulnerable and provide them with proof of indigent status to allow them to access all services included in the PBF benefit package free of charge.
b. Unit prices for services delivered to the targeted ultra-poor are adjusted to compensate for the loss of revenues that health facilities experience if not charging user fees. These unit prices were calculated to cover exclusively loss of income from user fees, not to provide any explicit additional incentive to providers to care for the ultra-poor. The additional payments to compensate for loss of revenues from user fees were tied to services normally offered against payment of direct user charges at point of use (e.g. curative consultations, delivery services). No additional compensation was added to the basic PBF case-based payments for services already provided free of charge to the general population (e.g. antenatal care, HIV and TB testing and treatment, vaccinations, growth monitoring).

iii. *T3*: Traditional PBF + systematic targeting and subsidization for the poor + provider motivation to offer services to the ultra-poor. T3 operates the same PBF contracts as T1 and T2 and includes the same provisions to care for the ultra-poor than T2. The core difference relates to the unit prices applied in T3, whereby services provided to the ultra-poor were reimbursed at a higher rate than in T2, approximately 120% to 150% of the expected cost of service delivery. This was meant to compensate for loss of income from user fees, while at the same time offering providers an additional financial incentive to provide services to the ultra-poor. In line with what described above, these additional payments only pertained to services normally offered against payment of direct user charges at point of use.

iv. *T4*: Traditional PBF + Community-based health insurance (CBHI), including targeting of the poor. In this case, PBF, applying the same contractual model as described above, was introduced in parallel to a CBHI. The insurance scheme was rolled out with support from the NGO ASMADE, building on the experience of a scheme that had already operated for several years in Burkina, while integrating elements of the model that the government had envisioned for the later Régime d'Assurance Maladie Universelle (RAMU). Insurance was offered to the entire population at a yearly premium of 3900 FCFA (~ 7 USD) per person. Targeting took place following procedures similar to the ones used in T2 and T3 areas and the premium for the ultra-poor was fully subsidized by the program. The insurance benefit package included a wide range of primary and secondary healthcare services. Payments to providers were made by both the insurance (in place of user fees) and by the PBF program, as case-based rewards at the same as in T1.

Across PBF interventions, adjustments to case-based payments further made according to the remoteness of the catchment population, staffing levels, and distance from the district capital, so that remote and disadvantaged facilities received higher case-based payments than easily accessible and better-equipped facilities. This approach resulted in the generation of nine different possible prices for the services included in the PBF benefit package (beyond the adjustments made in T2 and T3 to compensate for loss of income from user fees and to offer an additional financial incentive to provide care to the ultra-poor).

Across T2 and T3, SERSAP (Société d'Etudes et de Recherche en Santé Publique) was charged of conducting the identification of the ultra-poor using a community-based targeting approach. Targeting procedures have been described in detail in [10] and [11].

The PBF program was rolled out in six regions (Centre Nord, Centre Ouest, Nord, Sud Ouest, Boucle du Mouhoun, and Centre Est), purposely selected by the MoH and its development partners as having health indicators below the national median at the onset of the intervention. Within each region, the MoH purposely selected two districts to receive PBF on the basis of particularly poor outcomes on four key

indicators: (i) contraceptive prevalence rate; (ii) assisted deliveries; (iii) antenatal consultations; and (iv) post-natal consultations.

1.4. Study objective and research questions

The overall objective of the impact evaluation was to assess the impact of the PBF program on health service utilization and quality of service delivery across a wide range of targeted services. In line with what described above, the specific focus of this impact evaluation was on estimating the added benefit of combining PBF with equity measures.

The main research questions fitting the abovementioned objectives were:

1. What is the effect of the PBF program (irrespective of intervention package) on selected human resources, service quality, service utilization, and health status indicators, compared to status quo?
2. What is the effect of the different PBF design options on selected human resources, service quality, service utilization, and health status indicators, compared to status quo?
3. What is the added benefit of implementing T2, T3, and T4 compared to the standard T1 on selected human resources, service quality, service utilization, and health status indicators?

as well as across research questions 1-3:

4. What are the effects when considering only the most vulnerable segments of society, i.e. the ultra-poor?

2. Methods

2.1. Study design

The study design chosen resulted from an iterative discussion among the impact evaluation stakeholders, specifically the MoH, the World Bank, and the independent impact evaluation research team, in which policy interests were appraised against both scientific considerations and implementation concerns. While the primary interest of the technical partners, most specifically the World Bank, was to test the additional impact of moving from a simple PBF model to one that combined PBF with specific equity measures, the MoH of Burkina Faso was also interested in identifying the overall impact of introducing PBF, irrespective of its specific intervention modality, to substantiate the evidence produced by the pre-pilot.

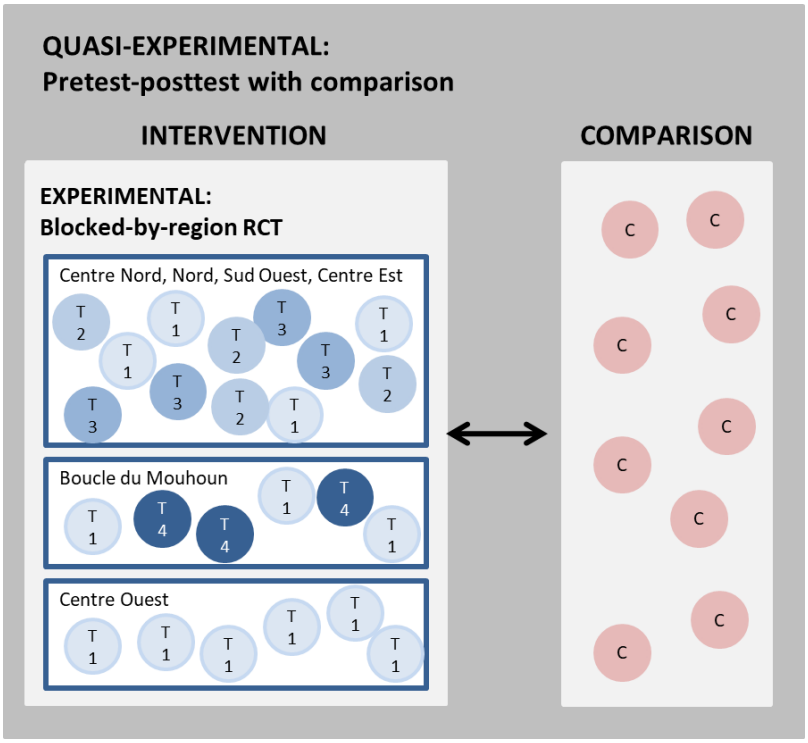
From a scientific point of view, the optimal design to accommodate both the policy-driven research interests put forward by the MoH and those put forward by the World Bank would have been to randomly allocate the four different PBF intervention packages described above across the selected districts and to include a random allocation to control (i.e. status quo) within the same districts. This design was judged to be unfeasible from the policy makers' perspective for three reasons. First, the implementation of a CBHI scheme appeared to be too complex an intervention to be allocated randomly across facilities in twelve districts. The level of know-how necessary to facilitate insurance implementation was absent in most districts, hence from the very onset of the discussions, the Government made clear its intention to test the insurance model in conjunction with PBF exclusively in the one region, Boucle du Mouhoun, where prior experience with insurance implementation was present. Second, it became apparent that the implementation of the targeting component would be quite costly and that funds may not suffice to carry out the ultra-poor selection process across all twelve concerned districts. Third, policy makers feared that randomizing facilities to intervention and control within a single district could have led to conflict arising as certain facilities and their respective communities would have been systematically excluded from the intervention, but would have known of it given geographical proximity. In addition to these policy concerns, by the time the Burkina impact evaluation was designed, experience from other settings had already revealed that the co-existence of intervention and control facilities within the same district could lead to extensive spillover effects, primarily due to consumers' mobility across facilities and to shared management by a single DHMT, posing a real challenge for effect identification at the analytical stage [12][13].

Against this background, the impact evaluation was design as a quasi-experimental design with a nested experimental component (**Figure 1**). In practice, this meant that for each intervention region with two intervention districts, two additional districts in the same or in a neighboring region were selected as controls, identified due to their relative proximity and similarity to the intervention districts in the targeted regions.

The twelve control districts received no intervention at all. The four PBF packages described earlier were implemented in the twelve concerned districts as follows:

a. In eight districts, the T1, T2, and T3 intervention packages were randomly allocated to health facilities and their catchment areas;
b. In two districts, the T1 and T4 intervention packages were randomly allocated to health facilities and their catchment areas;
c. In two districts, only the T1 intervention package was implemented.

Figure 1: Study design



In relation to the research questions presented in 1.4, the quasi-experimental element of the design was used to assess the overall impact of the PBF program (irrespective of the specific intervention package) vis à vis status quo health service provision. The experimental element of the design was used to assess the specific added benefit of introducing equity measures (as in T2, T3, and T4) alongside the implementation of standard PBF (T1).

2.2. Randomization

Within all intervention districts with randomization, facilities were randomized into the different intervention arms in randomization ceremonies, attended by all health facility in-charges, district health managers, and other important district and regional stakeholders to maximize transparency. The randomization procedure is described in detail in the baseline report [14]. In brief, health facility in-charges then took turns

drawing facility names from a box containing all health facility names in the district. Starting with a predefined intervention arm, facilities were then assigned to intervention arms in the order in which they were drawn from the box (i.e. 1st facility: T1, 2nd facility: T2, 3rd facility: T3, 4th facility: T1, etc.).

For pragmatic concerns, no randomized allocation was done at the level of the district hospitals. All twelve district hospitals in the intervention districts, as referral hospitals to all health facilities in the respective districts, were assigned to the T2 intervention packages. Further excluded from the randomization were those health facilities in the Nouna district in whose catchment areas a CBHI had already been active prior to the start of the pilot intervention.

Table 2 provides an overview over study districts and the intervention arms their facilities were assigned to. **Figure 2** shows the intervention and control districts and their health facilities by assigned intervention arm.

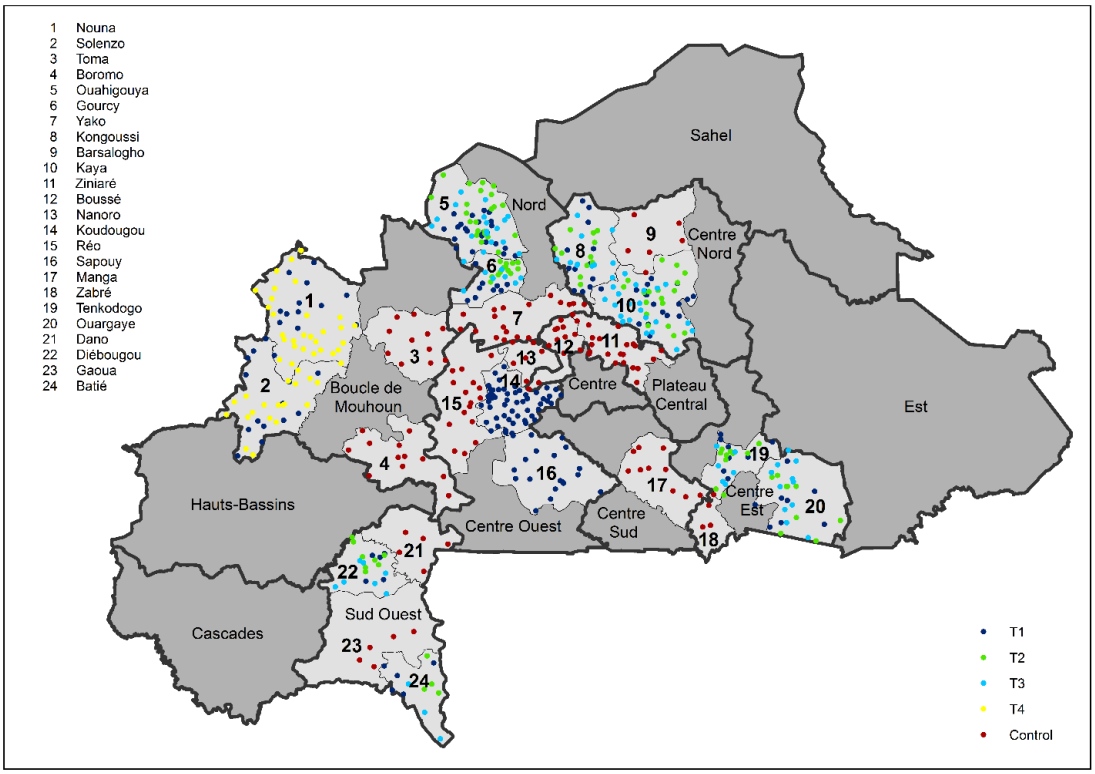
Table 2: Study regions and districts

Region	Intervention districts		Control districts[2]
Boucle du Mouhoun	Nouna	T1 (13), T4 (31)	Boromo
	Solenzo	T1 (16), T4 (17)	Toma
Centre Nord	Kongoussi	T1 (12), T2 (11), T3 (11)	Barsalogho
	Kaya	T1 (20), T2 (16), T3 (24)	Ziniaré
Centre Ouest	Koudougou	T1 (54)	Nanora
	Sapouy	T1 (18)	Reo
Sud Ouest	Gourcy	T1 (10), T2 (12), T3 (7)	Yako
	Ouahigouya	T1 (28), T2 (23), T3 (17)	Boussé
Centre Est	Batié	T1 (5), T2 (3), T3 (3)	Dano
	Diébougou	T1 (5), T2 (7), T3 (7)	Gaoua

Note: The arrangement of intervention and control districts next to each other does not imply that pairs of intervention and control districts are matched, but is done for economy of presentation. Numbers in brackets are the number of health facilities included in each intervention arm.

[2] Ziniaré, Boussé and Manga are part of a different region. This is because there are not a sufficient number of districts in the Nord, Centre-Nord and Centre-Est regions for having two control districts in the same Region as the two intervention districts. These two districts were identified due to their relative proximity and similarity to the intervention districts in the targeted regions.

Figure 2: Study districts and facilities



2.3. Indicator selection

A list of 31 indicators was selected for the impact analysis based on the following considerations and criteria.

a. **Theoretical relevance and alignment with the PBF theory of change.** We selected outcome indicators to reflect areas where PBF, as implemented in Burkina Faso, could have been expected to produce change. We included indicators at different levels, namely indicators related to human resources, to service quality (including perceived quality), to health service utilization and coverage, and to population health status. Our ambition was to document change attributable to PBF from the most immediate and expected (changes in human resource attitudes and behavior) to the ultimate objective of the program (changes in population health).
b. **PBF program indicators.** To the extent possible, we aimed to include indicators aligned with the primary care level indicators targeted by the PBF program. We focused on primary care indicators given that our analytical approach focused on estimating the impact of PBF among primary level facilities. In regards to indicators of quality of care, it is important to note that the indicators are not fully aligned with the quality indicators incentivized by PBF, which largely contain indicators related to availability of inputs and process indicators based

on document review. The impact evaluation, in contrast, relied on direct observation of actual care provided. For instance, PBF incentivizes correct use of the PCIME checklist as determined by document review, whereas our corresponding indicator pertains to adherence to PCIME as directly observed, irrespective of the checklist. While actual care provided ultimately is what the intervention aims at, it is important to remember this slight misalignment between what PBF purchased and the quality of care indicators used for the purpose of the impact evaluation.

c. **Alignment with national and international coverage and quality of care standards.** Service utilization and health outcome indicators were as closely aligned as possible with international key indicators. Quality of care indicators were as closely aligned as possible with national standards and treatment guidelines.
d. **Baseline figures.** We further considered baseline indicator levels in the selection and exact definition of indicators to allow for the detection of any either positive or negative change, should such a change have taken place. For instance, albeit present in many other similar evaluations, we excluded "Proportion of pregnant women with at least one antenatal care visit" given that the indicator was at 97% at baseline already and therefore we could have not easily detected an additional increase attributable to PBF. In such instances, we included a different yet similarly relevant indicator (in this specific case "at least four ANC visits") better suited to capture the same underlying change in service coverage and/or quality of service delivery.
e. **Data quality.** In the indicator selection and definition process, we took into account the data quality to ensure that the planned calculations would be feasible. Due to data quality issues at baseline (too many missing values), for instance, we were unable to include an indicator on staff absenteeism (although it would have been theoretically relevant), and due to SNIS data incompleteness, we were unable to include an indicator on the number of HIV tests performed.

As described in more detail in 2.5, we had some concerns regarding the endline sample of children for which immunization information, biomarkers, and anthropometric measurements were collected. We therefore assessed these indicators both with primary data as well as with corresponding SNIS data to the extent possible.

Table 3 contains the full list of indicators for which impact estimations were performed. Details on indicator definition and calculation are given in results.

Table 3: List of indicators selected for the impact evaluation

	Indicator	Data source
Indicators pertaining to human resources		
1	Proportion of staff having been evaluated for their performance in last year	Health worker survey
2	Health workers' perceived individual agency	
3	Health workers' satisfaction with the physical work environment	
4	Health workers' satisfaction with their compensation	
5	Health workers' satisfaction with management and supervision	
6	Health workers' intrinsic motivation	
Indicators pertaining to health service quality		
7	Proportion of facilities with permanent availability of power and safe water in the last 7 days	Health facility assessment
8	Proportion of facilities with at least one unit of 23 essential drugs in stock	
9	Proportion of observed ANC cases having received three key routine ANC services	Direct observation
10	Proportion of observed ANC cases having received patient education on three key elements	
11	Proportion of children observed in curative consultations having been assessed for all IMCI danger signs	Direct observation
12	Proportion of children observed in curative consultations having been assessed for common childhood illness symptoms according to IMCI	
13a	Proportion of ANC clients perceiving adequate quality of care on seven key elements	Exit interview
13b	Proportion of U5 consultation clients perceiving adequate quality of care on seven key elements	Exit interview
13c	Proportion of curative consultation clients aged 5 or older perceiving adequate quality of care on seven key elements	Exit interview
Indicators pertaining to the utilization of reproductive health care services		
14	Proportion of recently pregnant women with at least four ANC visits	Household survey
15	Proportion of recently pregnant women with an ANC visit within first four months of pregnancy	
16	Proportion of recently pregnant women having received at least 2 doses of tetanus vaccine during pregnancy	
17	Proportion of recently pregnant women having been offered HIV testing during pregnancy	
18	Number of HIV-positive mothers who have completed prophylactic ARV treatment	SNIS
19	Proportion of recently pregnant women who have delivered in a formal health facility	Household survey
20	Proportion of recently pregnant women with at least one PNC visit within 6 weeks after delivery	
21	Proportion of recently pregnant women with at least three PNC visits within 6 weeks after delivery	
22	Proportion of non-pregnant women aged 15-49 who use modern family planning methods	
Indicators pertaining to the utilization of preventive child health services		
23	Proportion of children aged 12-23 months who are fully immunized (primary data) Number of children aged 0-11 months fully immunized (SNIS)	Household survey, SNIS
24	Proportion of children aged 0-11 months who have participated in growth monitoring in last 6 months (primary data) Number of new growth monitoring visits of children aged 0-11 months (SNIS)	
25	Proportion of children aged 12-23 months who have participated in growth monitoring in last 6 months	Household survey
Indicators pertaining to the utilization of curative health care services		
26	Number of patients under age 5 having sought curative services	SNIS
27	Number of patients age 5 or older having sought curative services	
Indicators pertaining to population health status		
28	Proportion of children aged 0-59 months who are severely stunted	Household survey
29	Proportion of children aged 0-59 months with severe acute malnutrition	
30	Proportion of children aged 6-59 months with anemia	
31	Proportion of women aged 15-49 years with anemia	

2.4. Data sources

The impact evaluation relied on three main sources of data to measure the proposed indicator set.

1. A **household survey,** implemented at baseline (October 2013-March 2014) and endline (April-June 2017).
2. A **facility-based survey,** also implemented at baseline and endline, including different tools for data collection: a health facility assessment, a health worker survey, direct provider-patient observations, and patient exit interviews.
3. Data from the MoH's routine health information system (système national d'information sanitaire, **SNIS**).

For the household and facility-based survey, we used a slightly revised version of the data collection tool set included in the HRITF impact evaluation toolkit tailored to the needs of this specific impact evaluation and to the Burkinabé context (**Table 4**).

At endline, we added a short section on basic health worker perceptions and knowledge of the intervention to the health worker survey. We present these data descriptively in section 3.1.

Primary data collection was managed by the Centre MURAZ in Bobo-Dioulasso with support by the HIPH team. Details are available in data collection plans and reports prepared and submitted by Centre MURAZ. In line with the standard procedures for HRITF-funded PBF impact evaluations, facility-based survey teams spend one day per health facility to perform all assessments, interviews, and observations. Facility visits were arranged with facility in-charges in advance to ensure availability of staff members. At baseline, for this specific set of tools, data were collected on paper and digitalized at Centre MURAZ using a double entry strategy, while at endline, data were collected with tablets. Household survey teams also spend one day per village/facility catchment area. In order to ensure efficient data collection, the data collection team supervisors will spend part of their work time traveling ahead of their teams to observe social protocols and finalize sampling before arrival of the data collection team. Household survey data were collected electronically both at baseline and endline. Centre MURAZ staff supervised the data collection with support of HIPH staff members. An independent quality assurance mission was commissioned to the Institut de Recherche en Sciences de la Santé (IRSS) by the World Bank.

Independent of the primary impact evaluation commissioned to HIPH and Centre MURAZ by the World Bank and PADS, the HIPH team obtained ethical and MoH approval to use data from the routine health information system (SNIS) for additional research purposes. In the SNIS, monthly patient counts on health services are collected for each health facility. Since 2013, data is available in a web-based database. For the purpose of this report, we used data on a number of relevant indicators difficult to capture within the sampling strategy for the primary data collection (e.g. HIV, curative consultations), as well as for a robustness check on a few indicators for which we had concerns about potential sampling bias.

Table 4: Baseline and endline household and health facility surveys

Data collection tool	Respondents	Type	Survey Instrument	Description of Data
Household survey	Currently pregnant women; Women who have had a child in the 2 years preceding the survey	Quantitative	Adapted HRITF household survey instrument	Health service use, health care seeking behaviors and barriers to use for MCH services, health expenditures, perceptions of health service quality
Household survey	Currently pregnant women, non-pregnant women who have had a child in the 2 years preceding the survey, children under five	Anthropometry & biomarkers	Not applicable	Rapid diagnostic tests for malaria & anemia; Height and weight measurements
Facility assessment (inventory)	Facility in-charge	Quantitative	Adapted HRITF health facility questionnaire	Facility staffing, infrastructure, drugs supply, equipment, supervision, HMIS reporting and management, user charges, facility revenue
Health worker survey	Health care workers	Quantitative	Adapted HRITF health facility questionnaire	Staff training, compensation, motivation, satisfaction, and knowledge (incl. at endline knowledge and perceptions of PBF)
Direct patient-provider observation (Under-five & ANC)	ANC clients New under-5 patients for curative care	Quantitative	Adapted HRITF health facility questionnaire	Case management, treatment and counseling provided to patients.
Patient exit interviews	ANC clients New under-5 patients for curative care New over-5 patients for curative care	Quantitative	Adapted HRITF health facility questionnaire	Patient's (or caretaker's) perception of quality of care and satisfaction

2.5. Samples

This subsection contains information on the sampling strategies employed and samples obtained.

Health facility survey

The **facility survey** was conducted in all primary- and secondary-level health facilities in the intervention districts, as well as a random sample of health facilities in the control districts for an intervention-control facility ratio of approximately 3:1. This amounted to a total of 537 primary and 24 secondary facilities in the 12 intervention and 12 comparison districts surveyed at both baseline and at endline. The health facility sample was a fully balanced panel, accordingly. **Table 5** provides an overview over the health facility sample.

Table 5: Health facility sample

	All intervention	T1	T2	T3	T4	Control
Total	432	199	97	86	50	129
Level of care						
Primary	420 (97 %)	197 (99 %)	89 (92 %)	86 (100 %)	48 (96 %)	117 (91 %)
Secondary	12 (3 %)	2 (1 %)	8 (8 %)	0 (0 %)	2 (4 %)	12 (9 %)
Health facility type*						
Centre Medical	7 (2 %)	3 (2 %)	0 (0 %)	4 (5 %)	0 (0 %)	1 (1 %)
CSPS	399 (94 %)	188 (95 %)	85 (96 %)	79 (92 %)	47 (98 %)	116 (99 %)
Dispensary	7 (2 %)	4 (2 %)	2 (2 %)	1 (1 %)	0 (0 %)	0 (0 %)
Private	7 (2 %)	2 (1 %)	2 (2 %)	2 (2 %)	1 (2 %)	0 (0 %)
Location*						
Urban	43 (10 %)	24 (12 %)	7 (8 %)	8 (9 %)	4 (8 %)	2 (2 %)
Rural	377 (90 %)	173 (88 %)	82 (92 %)	78 (91 %)	44 (92 %)	115 (98 %)

* primary-level only, as secondary-level facilities were excluded from the analyses (see 2.6)

All clinical skilled health care available on the day of the interviewer team visit was to be interviewed with a structured **health worker survey**[3]. Due to high staff turnover rates as well as time and budget constraints, the health worker survey was implemented as a repeated cross-sectional survey rather than a panel. However, for endline participants, information was recorded allowing to link them to their respective baseline data, should they have happened to also have been interviewed at baseline, for a reconstruction of a partial panel. Only 14.5% of endline health workers had been interviewed at baseline. We therefore did not make use of the partial panel for this report.

Observations of antenatal care consultations were not performed at all facilities included in the sample as many primary care facilities do not offer ANC services on all days of the week. This was particularly an issue at baseline, whereas at endline, it appeared that many health facilities had changed their procedures to offer ANC more frequently. As the tight timelines and budget allowed for only one day per facility and interviewer teams were unable to align their field schedules fully with facility ANC schedules, ANC was observed at only 67% of facilities at baseline, and 94% of facilities at endline. Note that this was foreseen at the on-set of the study and does not introduce problems in terms of power/sample size. Further, we have no reason to believe that facilities without ANC observations are systematically different from facilities where ANC was observed. At each facility where ANC could be observed, the target was to observe five consultations.

Under-five patient-provider observations were performed at all sampled health facilities. Specifically, the target was to observed five consultations for children under 5 presenting with a new condition (i.e. no follow-up or routine visits) at each facility. At baseline, this sometimes proved challenging due to a lack of patients.

Exit interviews were performed with patients following their consultations for antenatal care, curative consultations for children under 5 (more specifically, their caregiver), and patients aged 5 and above who presented for curative services (or their caregiver if a child). For each service category, the target was five exit interviews at each facility. For patients under 5 and presenting for ANC, the aim was to interview those patients upon exit whose consultation had prior been observed. This was possible in approximately 95% of cases.

Table 6 provides an overview over sample sizes for the health worker survey, the direct observations, and the exit interviews at baseline and endline. The somewhat larger samples at endline are due to higher availability of patients to be observed/interviewed and small differences in the organization of data collection, notably higher interviewer effort levels at endline due to reorganization of supervisory structured and the external quality assurance activity (see 2.4).

[3] At secondary-level facilities, a random sample of three health workers with maternal and child health service delivery responsibilities was to be taken.

Table 6: Health worker survey, observation, and patient exit interview samples

	Baseline						Endline					
	Int total	T1	T2	T3	T4	Ctrl	Int total	T1	T2	T3	T4	Ctrl
Number of health workers surveyed	1076	535	209	193	139	285	1410	683	296	273	158	359
Number of observed ANC consultations	1337	647	295	270	125	200	1945	930	430	388	197	578
Number of observed cons. of children under 5	1687	763	370	378	176	359	2070	970	440	436	224	579
Number of exit interviews: ANC clients	1334	645	294	270	125	201	1973	936	422	398	217	581
Number of exit interviews: Children under 5	1683	763	369	374	177	359	2053	960	433	430	230	581
Number of exit interviews: Clients age 5+	1729	805	388	370	166	411	2005	954	421	421	209	565

Household survey

One village was randomly selected from the catchment area of each of the 523 public primary health care facilities[4]. Within each sampled village, at baseline, 15 households were randomly selected for interview among all households meeting the following main sampling criterion:

> Households with at least one currently pregnant woman or at least one woman who ended a pregnancy within the 24 months prior to the survey.

At endline, following discussion with World Bank staff and based on their experience in other settings, it was decided to change the originally envisioned repeated cross-sectional data collection strategy to the construction of a partial household panel.

In practice, at endline, the same villages as at baseline were sampled. Within each village, baseline households were first revisited. In case they still fulfilled the abovementioned sampling criterion, they were resampled and included in the endline survey. This was the case for 53% of baseline households. 38% of baseline households were no longer eligible, 5% of baseline households could not be found at endline, and 4% of baseline households were still eligible, but refused to participate again. Non-eligible, not found, or refusal households were replaced with another household from the same village randomly selected from a list of all not already included eligible

[4] Secondary-level and private health facilities do not have catchment areas in the Burkinabé system.

households, but meeting the same sampling criteria. In panel households, baseline unique individual identifiers were re-used at endline to allow linking across years.

Within each household, the following information was collected:

- Household demographic and socio-economic profile;
- Deaths over the prior 10 years;
- Individual-level illness reporting for both children and adults of chronic and acute illness and relative curative service use;
- Individual-level women of reproductive age: pregnancy and birth history, family planning use; and if pregnancy in the last two years: utilization of maternal health care services;
- Individual-level children under 5 years: immunization status and use of growth monitoring services;
- Individual-level women of reproductive age and children under 5: weight and height measures, anemia test, rapid diagnostic test for malaria.

Table 7 contains sample sizes at household level as well as for the individual household member subsamples relevant to the impact evaluation indicators.

Note that the creation of the partial panel did introduce a systematic bias in that not surprisingly, fertility tended to be slightly higher among women in panel households

Table 7: Household survey sample

	Baseline						Endline					
	Int total	**T1**	**T2**	**T3**	**T4**	**Ctrl**	**Int total**	**T1**	**T2**	**T3**	**T4**	**Ctrl**
Number of households	6224	2896	1274	1320	734	1754	6204	2889	1270	1310	735	1753
Number of women age 15-49 surveyed	7766	3725	1602	1606	833	2233	8432	4030	1722	1752	928	2578
of which with a pregnancy in last 24 months	5074	2497	970	1026	581	1494	4932	2334	983	1010	605	1439
of which measured (anthrop., biom.)	6689	3177	1386	1422	704	1286	7860	3688	1676	1655	841	2335
Number of children under 5 surveyed	9230	4512	1831	1842	1045	2623	10851	5183	2210	2179	1279	3160
of which measured (anthrop., biom.)	8209	4106	1591	1571	941	2185	10170	4779	2143	2069	1179	2941

(average parity of 4.4, compared to 4.2 in the overall sample at endline)[5]. However, this was constant across intervention and control areas, so that the DID estimates should not have been affected, and we did not find substantial differences on other pertinent variables.

In regards to the children under 5 surveyed and measured, we noticed some discrepancies calling into question the representativeness of the samples. Specifically, we observed a strong increase in children under 5 in the sampled households (+42%) as indicated in the household-related part of the household survey, which cannot be attributed to population growth alone. At the same time, samples for children surveyed and measured are of similar sizes at baseline and endline (**Table 7**). Two scenarios are thinkable, a) that at baseline, not all children were registered in the initial listing of household members, or b) that not all children under 5 who should have been surveyed/measured at endline were in fact surveyed and measured. An inspection of the data combined with discussions with the data collection agency leads us to believe that a combination of both factors was responsible. Additional analyses show that there was some bias in regards to demographic characteristics of the surveyed/measured children in comparison to the full sample, although we do not have much data to test this, which was not always fully consistent across intervention and control groups. More detailed information Is provided in **Appendix A**.

The issue pertains primarily to impact indicators 23-25 (preventive child health service utilization) and indicators 28-30 (health status indicators). Unfortunately, beyond controlling for demographic characteristics to the extent possible and interpreting estimates with caution, we are limited in the extent to which we can address the issue directly. As indicated in 2.6, we repeated all analyses on the panel subsample, with results supporting the overall impact estimates. We also duplicated the preventive child health service utilization indicators with routine data from the SNIS to the extent possible. Results also support those of the main analyses. For the population health indicators, however, such data are unfortunately not available. All available information considered together, we recommend some caution in the interpretation particularly of indicators 28-30, pertaining to children's health status, but have no strong reason to believe that potential biases have impacted the evaluation results to an extent that the overall 'story' is affected.

SNIS

We extracted data for all facilities included in the facility survey. The data reflect monthly patient counts for different services categories for each facility. For indicators based on SNIS data, we used data for the following time intervals to generate a baseline and an endline estimate, respectively:

- Baseline: October 2013 – March 2014
- Endline: October 2016 – March 2017

[5] In 2010 (latest DHS), data on pregnancy spacing among women with more than one child in the last 5 years indicated that 73 % had the second child within 3 years (i.e. the time lag between our baseline and endline) after the first (http://reprolineplus.org/system/files/resources/Burkina-PPFP.pdf)

The periods were selected to reflect the data collection strategy applied in the context of other PBF impact evaluations through means of questions included in the facility survey. Accordingly, we then computed the total across the respective 6 months as the baseline and endline indicator value for each facility.

2.6. Analysis

In alignment with the standard HRITF analysis strategy for the PBF impact evaluations, we used a Difference-in-Differences (DID) approach for all impact analyses.

Overall model specifications

In alignment with the quasi-experimental intervention design with a nested experimental component and in response to the three main research questions (1.4), we performed three types of impact analyses for each indicator:

1) Step 1: Overall PBF effect compared to status quo:
   In order to address the research question "What is the impact of PBF compared to status quo?", we relied on the quasi-experimental design, comparing PBF facilities (pooled across all intervention arms) to control facilities.

2) Step 2: Disaggregated PBF effects compared to status quo
   In order to address the research question "What is the impact of different PBF design options compared to status quo?", we also relied on the quasi-experimental design, comparing each intervention arm to control facilities.

3) Step 3: Disaggregated effects of T2, T3, and T4 compared to T1
   In order to address the research question "What is the relative added benefit of T2, T3, and T4 in comparison to the standard T1?", we relied on the experiment (i.e. randomized controlled trial) nested within the overall quasi-experimental design, comparing T2-T4 facilities to T1 facilities without making use of the control facilities.

This entailed number of separate regressions for each of the indicators: one regression for step 1, one regression for step 2, and two regressions for step 3 (one for those districts with T1, T2, and T3, and one for the two districts in Boucle du Mouhoun with T1 and T4).

In Step 1, the following regression equation will be estimated ('specification 1'):

$$Y_{dfit} = \alpha_f + \beta \cdot 2017_t + \delta \cdot [PBF_d * 2017_t] + \phi \cdot X_{it} + \epsilon_{dfit},$$

where Y_{dfit} is the outcome variable for individual *i* from facility/catchment area *f* in district *d* at time *t* with *t={2013, 2017}*. 2017_t is a dummy variable indicating endline

observations, thus coefficient β gives the time fixed effect. PBF_d is a dummy variable that equals one for individuals in PBF districts and zero for individuals in control districts. α_f are facility/catchment area fixed effects to capture time-invariant unobserved differences across health facilities, and X_{it} is a set of additional individual-level covariates (**Table 8**). Note that by applying facility fixed effects, the model implicitly captures unobserved baseline differences between treatment and control districts. Since the intervention was implemented at the district level, neither facility fixed effects nor our set of covariates should have affected the magnitude of the DID estimate, but only enhanced its precision. ϵ_{dfit} is the error term. Following common practice, standard errors were clustered at the district level, which is the level of treatment assignment for the quasi-experimental component of the study design [15]. The coefficient δ gives the DID estimate for the effect of being located in a PBF district when compared to non-PBF districts.

In Step 2, the following regression equation will be estimated ('specification 2'):

$$Y_{dfit} = \alpha_f + \beta \cdot 2017_t + \sum_{k=1}^{4} \delta_k \cdot \left[PBF_d^k * 2017_t\right] + \phi \cdot X_{it} + \epsilon_{dfit},$$

where PBF_d^k is are dummy variables that equal one for individuals from facilities/catchment areas in intervention arm PBFk, where k={T1,T2,T3,T4}. Individuals from facilities/catchment areas in control districts provide the comparison group. The DID estimates δ_k give the effects of PBFk in comparison to status quo (control districts). The remaining equation components are equal to specification 1. Note that as in specification 1, standard errors were clustered at district level as the level of treatment assignment in the comparison of each intervention arm with controls remains at the district level.

In Step 3, the following regression equations will be estimated ('specification 3'):

$$\mathrm{Y}_{fit} = \alpha_f + \beta \cdot 2017_t + \delta_2 \cdot [T2_f * 2017_t] + \delta_3 \cdot [T3_f * 2017_t] + \phi \cdot X_{it} + \epsilon_{fit}$$
$$\mathrm{Y}_{fit} = \alpha_f + \beta \cdot 2017_t + \delta_4 \cdot [T4_f * 2017_t] + \phi \cdot X_{it} + \epsilon_{fit}$$

where Y_{fit} is the outcome variable for individual *i* from facility/catchment area *f* at time *t* with *t={2013, 2017}* in the intervention districts. In step 3 models, standard errors were clustered at health facility/catchment area level, the level at which random assignment into intervention arms took place in the experimental component of the study design. The upper equation was applied to the eight districts where randomization over three intervention arms (T1-T3) took place. δ_2 and δ_3 give the DID estimate for the effect of being located in T2 compared to T1 and T3 compared to T1, respectively. The lower equation applies to Boucle de Mouhoun, where health facilities were randomly allocated into the T1 or T4 treatment[6]. Hence, δ_4 gives the DID estimate for the effect of being located in T4 compared to T1.

[6] Note that while we included data from all health facilities in catchment areas in steps 1 and 2, non-randomized facilities in the Nouna district (see 2.2) were excluded from step 3 models.

Table 8: Individual characteristics used as additional covariates

Indicator category	Control variables included in all models
Human resources	• Health worker characteristics: Type, age, sex seniority
Health service quality: ANC	• Health worker characteristics (see human resources) • Patient characteristics: Age, parity, literacy, marital status, distance household-health facility, SES
Health service quality: Children under 5	• Health worker characteristics (see human resources) • Patient characteristics: Age, sex, distance household-health facility, SES • Principal guardian characteristics: sex, literacy, marital status
Health service utilization and health status – women of reproductive age	• Individual characteristics: Age, parity, literacy, marital status, distance household-health facility, SES
Health service utilization and health status – children under 5	• Individual characteristics: Age, sex, distance household-health facility, SES • Principal guardian characteristics: Age, parity, literacy, marital status, current pregnancy

Key DID assumptions

Parallel trend assumption. The key assumption in using difference-in-differences, specifically in using the controls to approximate the counterfactual, is that control and intervention units would have in fact developed in the same manner in the absence of the treatment. This assumption is technically untestable, but inspection of pre-intervention trends allows for reasonable certainty of parallel trends. We were able to do so for a number of key health service utilization indicators for which equivalent SNIS data was available. Plots starting from January 2013 are provided in **Appendix B** and show pre-intervention trends reassuring of non-violation of the parallel trend assumption. For human resources, quality of care, and population health indicators, unfortunately, no routine information to test the parallel trend assumption is available.

DID does not necessarily require equal baseline levels in the intervention and control group or the different treatment groups to be compared, as such differences in absolute levels are factored out when taking the double difference. However, the absence of baseline differences further strengthens the validity of the assumption of parallel trend. Results of tests of baseline differences are provided in **Appendix C**. Significant baseline differences existed for almost all indicators when comparing all intervention arms to control (specification 1), for most maternal care indicators when comparing the different intervention arms to controls (specification 2), for selected indicators in the comparison of T2 and T3 to T1 (specification 3), despite

randomization, and for only three indicators when comparing T4 to T1 (specification 3).

Stable unit treatment value assumption. DID further assumes that treatment/non-treatment or the different treatment arms are clearly distinct, in the sense that a) each unit of observation is clearly in one state or the other, and that b) treatment is further uniform across all units of observations assigned to it. In regards to a), narrative evidence from implementers as well as the results of a parallel process evaluation led by the University of Montreal underline that the assumption is not fully valid. Although controls are separated from intervention in that they are in different districts, regional health managers have reported that there was spillover in that there was exchange among district health officers and competition among districts, leading to increased efforts even in control districts attributable to PBF. In regards to b), the evaluation would have ideally required that PBF was the only intervention implemented in the intervention districts, and that no interventions were implemented in the control districts. In a context in which not only the government continues to implement changes to improve access and quality of care, but where a multitude of donors and non-governmental organizations are active, this was impossible to achieve. Not only was the gratuité policy implemented nation-wide in June 2016 as discussed in the introduction, but a variety of other interventions pertaining to reproductive and child health was on-going in both intervention and control districts in parallel to PBF (**Appendix D**). Effect estimates therefore likely do not only reflect the pure impact of PBF, but also at least to some extent the concurrent implementation of PBF with other interventions with similar objectives. This is in particular true for the effect estimates pertaining to the impact of PBF compared to status quo (specifications 1 and 2).

Further analytical considerations

Exclusion of secondary-level facilities. In line with the World Bank focus on PBF in primary care, the mandate for the impact evaluation, and the following reasons, the 24 second-level facilities (CMA, CHR) were excluded from all analyses for this report. Results only pertain to primary-level care, accordingly.

- Hospitals do not directly serve first-level catchment areas since they function as referral centers for all primary care facilities within their respective second-level catchment region. We do not have corresponding household-level data for hospitals, accordingly.
- The data collection tools, research questions, and indicators proposed for the impact evaluation are focused on primary care services which are not supposed to be provided at hospital level. Inclusion of indicators corresponding to the hospital-level PBF indicators is largely impossible as no corresponding data were collected.

Exploiting panel structure versus treating data as repeated cross-sectional measurements. We have a balanced panel only at the health facility/enumeration section level. At the household level, only about half of baseline households were

eligible for participation in the endline survey (see 2.5). Similarly, whereas we have a complete health facility panel, re-interview rates at health worker level are only 14.5%, and there is no panel for patient-provider observations and exit interviews.

In light of this, we did not make use of the partial panels for the primary analyses reported in this report, acknowledging the likely existence of strong attrition bias and issues pertaining to sample size. We did capitalize on the panel structure at health facility/enumeration section level by including facility/enumeration section fixed effects into all DID models as described above. We also performed additional analysis using the partial panels at individual household member level as robustness analyses (see below).

Focus on population averages versus individual indigent/insurance status. In line with the objective of the impact evaluation commissioned to us, i.e. evaluate the impact of shift in financing policy, and with the study design, we used an Intent-to-Treat (ITT) approach to data analysis, looking at changes measured at the population level (e.g. how has health service utilization changed in T1 zones compared to T2 zones?) and not at the individual behavior (e.g. how has holding an insurance card/indigent card changed behavior in health service utilization?). The latter would be interesting but is not possible in the framework of the standard World Bank methodology for PBF impact evaluations.

The study design is not suited to rigorously evaluate PBF effects on healthcare seeking behavior specifically of targeted indigents or insured individuals, as

- we do not have a counterfactual, since targeting happened/insurance was put in place only in the respective intervention areas, but not in T1 or control areas
- the household sampling strategy and target sample size was not designed to capture a sufficiently large sample of targeted indigents/insured individuals for a meaningful and robust sub-group analysis, particularly for the respective indicators of interest, which pertain to sub-populations not frequently targeted [16]
- we have no data on (future) indigent targeting/insurance status at baseline except for panel individuals

Stratified analyses by socio-economic status. In response to research question 4, for indicators referring to health status and health service utilization (except those based on SNIS data), we further produced overall impact estimates as well as separate analyses pertaining to the ultra-poor, which, following discussions with the World Bank ream, we operationalized as the poorest 20% of households and their members. We chose 20% in line with common definitions (i.e. poorest population quintile) and the original target set by the World Bank for the indigent targeting process [16]. To identify the poorest 20%, we calculated a wealth index using the Standard Multiple Component Analysis (MCA) method. The following was used to create the wealth index: housing (type of building, number of rooms, water and energy supply), assets (TV, radio, fridge etc), house and fields owned, animals. Rural and urban areas were taken into account by choosing relevant assets in each area.

Composite quality of care indicators. All quality of care indicators are composite indicators, combining multiple elements to reflect the complexity of service delivery. While we present full DID results only for the overall composite indicators, we also performed analyses on the different individual elements of these composite indicators to understand which elements drive or hinder potential changes. These analyses are not systematically discussed in the text or displayed in tables so as not to overwhelm report readers. Where results were of interest, we highlighted them in the text, however.

Approach to missing values. Missing value rates on outcome variables were generally very low. We normally included observations into analyses only if all relevant information for the calculation of the respective outcome variable was available. This is with the exception of the psychometric indicators 2-6, the composite quality of care indicators 7-13c, and the SNIS-based indicators, where we calculated composite outcome variable values even in the case of missings so long as data were missing for no more than 30% of the subcomponents of an indicator. Specifically, for the psychometric variables, we calculated the mean over all respective items. In case of missings on 30% or less of the items, the mean was calculated over the remaining items. In case of missings on more than 30% of the items, the health worker was not considered in the analysis. For the composite quality of care indicators, in case of missings on 30% or less of the subcomponents, the indicator was calculated 'ignoring' the items with missing values, thereby implicitly assuming they were present or done. For SNIS data, missings were replaced with the mean of the other months in the respective 6-month interval if data were available for at least 4 out of the 6 months. If more than 2 months were missing in a respective 6-month interval, no value was calculated for the respective interval and facility.

We also did not exclude individuals based on missing in the covariates. This was the case for a small proportion of the sample due to minor sample misalignments (e.g. observed patients not interviewed upon exit, measured child omitted in household member listing and therefore without demographic characteristics) or mistakes in identifiers. In light of our observation in the analytical process that the addition of individual-level control variables over and above clustering and facility/catchment area fixed effects did not substantially change results and the overall small proportion of missings in demographic characteristics, we decided to impute missings in control variables rather than to exclude such cases. Specifically, we imputed using means or modes for other respondents at the same data collection time point, district, intervention arm, and with the same core demographic characteristics.

Robustness analyses

Inference issues related to the low number of clusters in model specifications 1 and 2. By design, the quasi-experimental part of the study (steps 1 and 2) is challenged by a relatively low number of clusters (24 districts). Too few clusters might lead to the estimation of downward biased standard errors and, consequently, to an over-rejection of the H_0 hypothesis that there is no program effect. Thus, there is an

elevated risk of postulating significant program effects when there is actually no effect detectable with our design. There appears to be no consensus in the literature yet as to which number of clusters is sufficient, but 24 clusters are on the lower end of the spectrum of sufficiency in available simulation studies (e.g. [17][18]). Further, studies have shown that the implications of too few clusters are considerably worse when clusters are strongly imbalanced in terms of within-cluster sample sizes as unfortunately is the case in our study design (e.g. [19][20]).

The available literature proposes some robustness tests [22][18], and shows that bias adjustments of cluster-robust standard errors can make quite a difference [22]. In an important simulation study, Cameron, Gelbach and Miller [18] investigated different recently suggested bootstrapping methods to obtain asymptotic refinement in a scenario with as few as five clusters. They found that the 'wild bootstrap' can considerably improve statistical inference of the coefficient estimate and produces much lower over-rejection rates of the H_0 than, for instance, the usual way of directly bootstrapping standard errors.

In line with this literature, we therefore applied the 'wild bootstrap' to all specification 1 models for a robustness check. In contrast to bootstrapping standard errors, the 'wild bootstrap' involves a bootstrap t-procedure [23], where the Wald statistic (already including cluster-robust standard errors as obtained from the `cluster' command in Stata) is bootstrapped, and where the resulting distribution of the Wald statistic values are used to form inference on the original Wald statistic obtained in our DID regressions.

For implementation in Stata, we used the (user-written) 'boottest' package. It produces a 95% confidence interval, based on a bootstrap t-procedure under the H_0 that the DID estimate obtained in our regression is true (i.e. the t-statistic from the H_0-hypothesis that the coefficient estimate actually equals $\hat{\beta}$ is bootstrapped). If the DID estimate lies within the interval (i.e. we cannot reject the hypothesis that true-coefficient equals $\hat{\beta}$), no concern about incorrect statistical inference due to too few clusters is warrented.

Estimates were all within the confidence intervals produced by the 'wild bootstrap' procedure, meaning that there is no risk of interpreting intervention effects where in fact there are none.

Household- and individual-level partial panels. As described in 2.5, a partial panel was generated at household and individual level. This partial panel contained only roughly half of the households at baseline and endline, so that we did not use it for the overall impact analyses. However, we did make use of the partial panel for additional robustness analyses.

Specifically, for indicators pertaining to women of reproductive age (14-22, 31), we repeated all analyses the subsample of women interviewed both at baseline and endline (also around 50% of the total sample), using individual fixed effects instead of

facility/catchment area fixed effects to improve estimates. Results largely support those of the overall impact analysis.

For indicators pertaining to children under 5 (23-25, 28-30), using the individual panel was not possible or sensible as children had grown out of the age tranche between baseline and endline. Instead, we repeated all analyses on the subsample of panel households, using household fixed effects instead of facility/catchment area fixed effects to improve estimates. Again, results largely support those of the overall impact analysis.

Potential bias due to the gratuité policy. As mentioned in the introduction, in June 2016, a user fee exemption policy (gratuité) targeting pregnant and lactating women and children under five came into effect. This policy was introduced national-wide and should have therefore equally affected intervention and control areas, thereby likely not distorting our DID estimates. However, we cannot exclude a saturation effect in the sense that utilization rates in both control and intervention districts increased close to the maximum, thereby masking potential intervention effects. For all SNIS indicators, the only ones for which this robustness analysis was possible, we repeated all analyses, using as endline data from one year prior to the main analysis and therefore before the start of the gratuité policy, so October 2015-March 2016 instead of October 2016-March 2017. Again, results largely support those of the overall impact analyses.

Quality of care: "All-or-nothing" indicators vs scores. All quality of care indicators are composite indicators, combining multiple elements to reflect the complexity of service delivery. We decided to take an "all-or-nothing" approach to calculating composites, meaning that a specific facility or observation was only considered of adequate quality if all sub-elements pertaining to the indicator were present or done. If one or several were not present or done, the case was considered of inadequate quality. Similarly, we only considered a patient to have perceived adequate quality of care if this was the case for all sub-elements. An alternative approach would have been to calculate scores reflecting the degree to which care was of high quality from worst possible to best possible. The advantage of this approach is that it gives a clearer idea of where facilities stand in relation to absolute quality standards, but this comes at the disadvantage of necessitating the application of weights to the sub-components in the calculation of the composite, which in the absence of clear clinical evidence is somewhat arbitrary. We decided to calculate scores in addition to the primary "all-or-nothing" indicators as an additional robustness check. In calculating the scores, we took the simplest approach of giving all respective subcomponents equal weight, but acknowledge that this is likely not reflective of their actual weights towards quality care and patient outcomes. We performed all analyses on both the "all-or-nothing" indicators and the scores. With a few exceptions, which we highlight in the results, results obtained with the score indicators support those obtained with the "all-or-nothing" indicators.

3. Results

In this chapter, we present key findings of the study, starting with descriptive findings related to health workers' perceptions and knowledge of the intervention at endline. In 3.2, we give an overview over how impact analysis results are presented and interpreted, followed by the presentation of the actual results of the impact analyses following in sections 3.3 – 3.8. For the latter, details can be found in **Appendix F**.

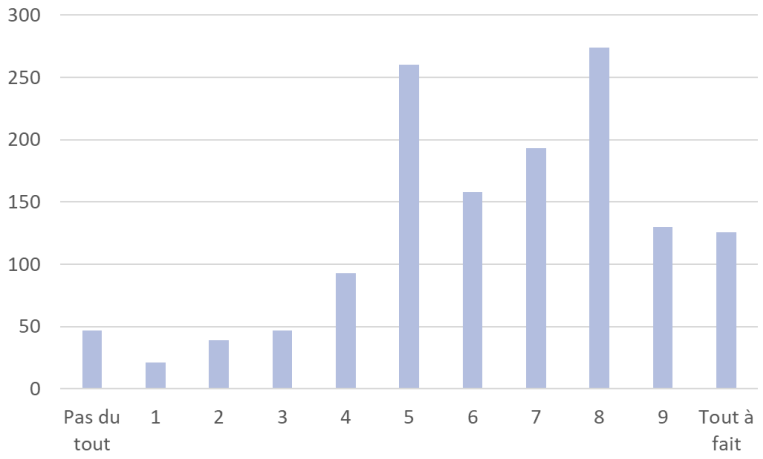
3.1. Health workers' perceptions and knowledge of the intervention

Data for the results presented in this section were collected with an additional section added to the health worker survey at endline for health workers in intervention facilities.

Overall satisfaction with PBF

As shown in the distribution of answers in **Figure 3**, there was large variation in health workers' overall satisfaction with the intervention.

Figure 3: Overall satisfaction ("Overall, how satisfied are you with the PBF?")

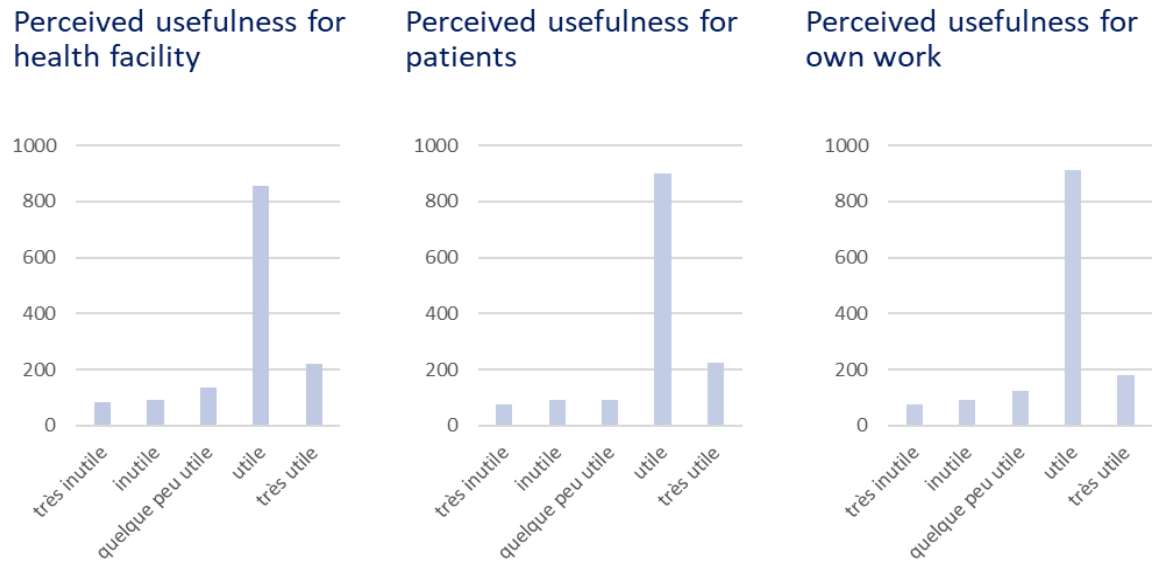


48% of respondents indicated wanting the project to continue as before or with minor modifications. 45% also indicated a wish for the project to continue, but with more fundamental adjustments, such as fundamental changes in the indicators, in the verification system, in the *outil d'indice*, or in the indigent selection process, or the abolition of the different intervention arms. 7% of respondents preferred the project not to continue.

Perceived usefulness of PBF

The vast majority of respondents found the project useful (although not very useful) for their health facility, their patients, and themselves (**Figure 4**). Only few did not perceive it as useful.

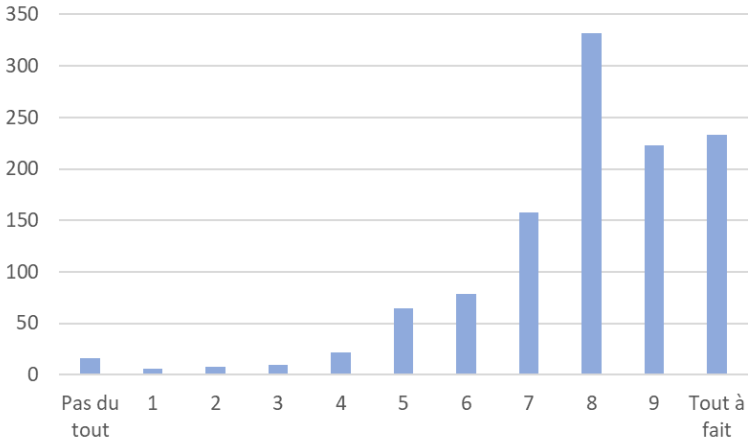
Figure 4: Perceived usefulness of PBF ("Do you find the project useful or not useful for … ")



Knowledge and perceptions regarding the performance quality evaluation

76% of respondents correctly[7] recalled the result of their facility's last performance quality evaluation determining whether the facility received or did not receive the quality bonus. The vast majority regarded the result of this last verification as fair in relation to what they perceived their facility's performance to have been, as shown in **Figure 5**.

Figure 5: Perceived fairness of the performance quality evaluation ("Was the last quality evaluation fair or unfair in light of your facility's performance?")

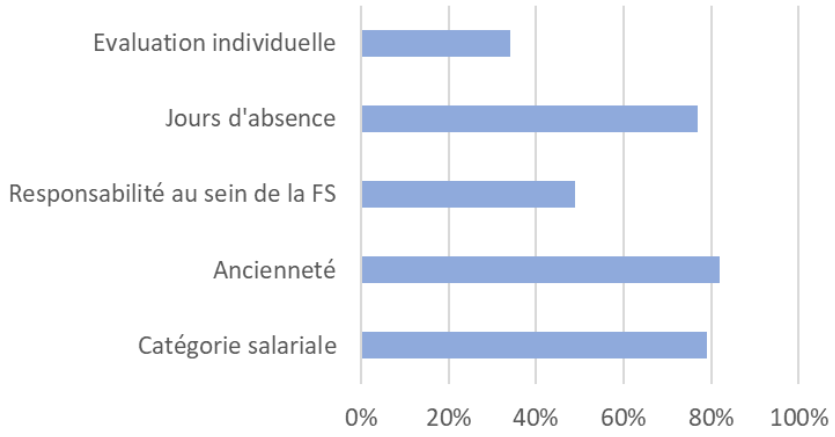


[7] Allowing for an error margin of +/- 5 points on the 0-100 quality score

Knowledge and perceptions regarding the individual financial incentive component

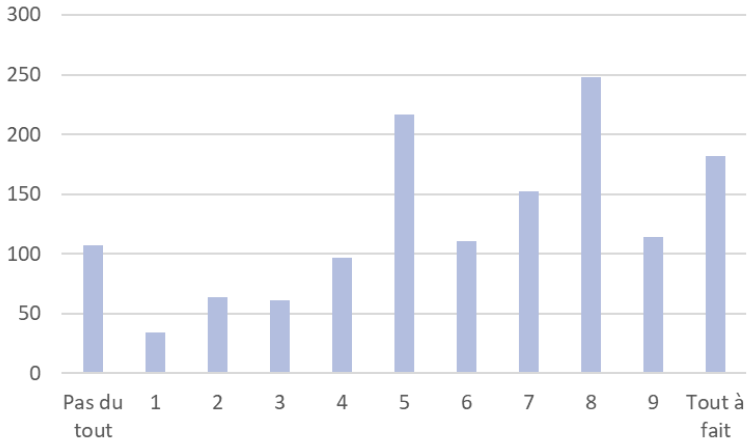
72% of respondents reported that they receive individual incentives in the context of PBF, whereas 28% indicated that they do not, despite working in a PBF facility.
Respondents were asked to list the criteria according to which PBF incentives are distributed among facility staff. As **Figure 6** shows, a high proportion of respondents correctly recalled days of absence, seniority, and salary category. Responsibility at the health facility (i.e. whether the health worker is in charge of the health facility or a department) or individual performance, in contrast, were not mentioned as frequently, despite also being criteria which are supposed to enter the calculation of individual amounts received by each staff member.

Figure 6: Proportion of respondents having recalled the five incentive distribution criteria



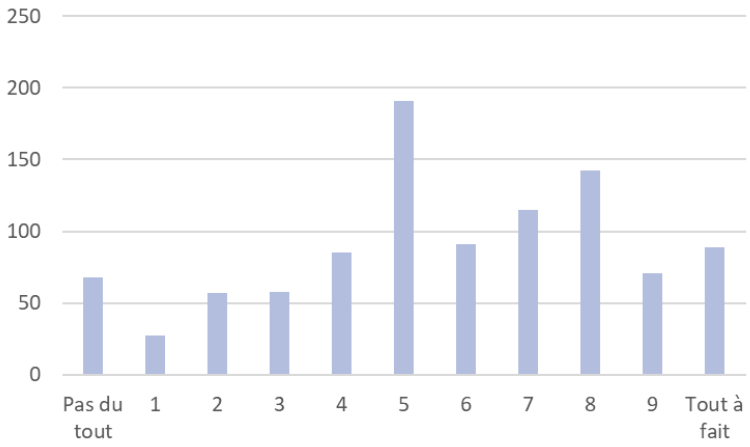
There was large variation in the sample in regards to whether they perceived this distribution mode as fair (**Figure 7**).

Figure 7: Perceived fairness of the individual incentive distribution mode ("Do you find the incentive distribution mode fair or unfair?")



Among those who reported to receive incentives, satisfaction with those incentives varied substantially as shown in **Figure 8**.

Figure 8: Satisfaction with the individual financial incentives ("Are you satisfied with the PBF "primes" that you receive?")



3.2. Presentation and interpretation of impact evaluation results

This section contains a brief description of how the results of the impact analyses are presented and interpreted in the following.

We will present results indicator by indicator, grouped by category as in **Table 3**. At the end of each subsection, we provide an overview table summarizing all results pertaining to a category. We then appraise findings globally in light of intervention aims, design, implementation, as well as experiences from other countries in the discussion section.

For each indicator, results are presented in a box. Each results box contains

- on the **left**, results pertaining to **specification 1**, the comparison of PBF to status quo (quasi-experiment): a graphic display of the development of intervention and control averages over time, with the corresponding DID estimate(s) below. For indicators where a stratified analysis on the poorest 20% was performed, two DID estimates are given, one for the overall sample ('all'), and one for the poorest 20% ('poor').
- at the **bottom**, results pertaining to **specification 2**, the comparison of the different intervention arms to status quo (quasi-experiment): DID estimates for the respective intervention arms. No graphs are provided.
- on the **right**, results pertaining to **specification 3**, the comparison of T2, T3, and T4 over and above the standard T1: graphic displays of the development of the different intervention arm averages over time and the corresponding DID estimate(s) below.

In addition to the results box, for each indicator, a short text describing how the indicator was defined and calculated and a summary of main results is provided. We further highlighted additional information if pertinent and interesting, for instance information on composite indicator subcomponents that drive the overall effect or on robustness analyses if we detected differences between the main and the sensitivity analysis. More detailed summary statistics and model details are provided in **Appendix F**.

Note that the means/proportions displayed in the **graphs** pertain to the overall sample and are not adjusted for district, facility/catchment area, or other characteristics. They might therefore suggest slightly different impact levels than the model-adjusted DID estimates. Not to overload the report, no separate graphs for the sub-sample of the poorest 20% are provided.

Important to note is also that **effect estimates** represent absolute change, i.e. absolute differences attributable to the intervention, not relative change from baseline values. As most indicators are binary, meaning that an individual either used or did not used a certain service, or that an observed consultation was either of adequate quality or not, for instance. Effect estimates for binary variables can therefore be translated into **percentage point (pp) change** attributable to the intervention. Exceptions are the psychometric indicators 2-6, where estimates are to be interpreted as intervention-attributable absolute point change on the 0-10 scale, and SNIS indicators, where effect estimates pertain to absolute intervention-attributable change in patient numbers.

Statistical significance of effect estimates is indicated with the common 'star notation', where * indicates statistical significance at the 10% level, ** at the 5% level, and *** at the 1% level. In 2.6, we discussed the issue of the low number of clusters for specification 1 and 2 models in relation to a risk of over-rejection of the H_0, i.e. of postulating intervention effects where there are in fact none. Results of the 'wild bootstrap' analyses ascertained that the statistically significant intervention effects found are real rather than resulting from the low number of clusters for all indicators except perceived quality of antenatal care. At the same time, the wild bootstrap confidence intervals for all statistically significant intervention effects except for satisfaction with the physical working environment and facility-based delivery (where 0 is the lower bound) contain 0. Results should be interpreted with the respective caution, accordingly.

However, the low number of clusters also has implications for the design's ability to detect real intervention effects compared to status quo as different from zero, i.e. its **statistical power**. Effectively, even large effect sizes are unlikely to reach statistical significance due to large confidence intervals due to the small number of clusters in the quasi-experimental component of the design. Note that this concerns only tests for whether estimates impact coefficients are significantly different from zero, not the validity of the estimate as such. This issue had been discussed in the planning stage at the study. At the time, the focus of the study had been on the nested cluster-randomized trial component of the study, where the level of treatment assignment is at facility level, and where power calculations had shown the number of clusters to be adequate [14] (see reproduction in **Appendix E**), and low power to detect effects in the quasi-experimental component of the study design was accepted.

In the following presentation and interpretation of results, we dealt with this issue as follows: For results pertaining to specification 1 and 2, i.e. the comparison of PBF to status quo in the quasi-experimental component of the study, we inspected effect sizes and interpreted in case of striking magnitude, even if they did not reach statistical significance. In results pertaining to the well-powered experimental part of the design (specification 3), in contrast, we did not interpret effect estimates that did not reach statistical significance as potentially different from zero.

3.3. Impact of PBF on human resources factors

In this section, results pertaining to the impact of PBF on issues related to human resources are presented. The specific indicators include:

1. Proportion of staff having been evaluated for their performance in last year
2. Health workers' perceived individual agency
3. Health workers' satisfaction with the physical work environment
4. Health workers' satisfaction with their compensation
5. Health workers' satisfaction with management and supervision
6. Health workers' intrinsic motivation

Data for all indicators in this section were collected using the individual health worker questionnaire (F2), administered to skilled clinical personnel sampled as described in section 2.5. For indicators 2-6, data was collected with psychometric scales, meaning that respondents answered to questions or statement on a response scale of 0-10, with 10 indicating high agency, satisfaction, or intrinsic motivation. Effect estimates indicate absolute point change on this 0-10 scale attributable to PBF compared to status quo (specifications 1 and 2), or of intervention arms T2-T4 over and above T1 (specification 3). Indicator 1, in contrast, is binary, meaning that each health worker in the sample was either evaluated or not. The effect estimates can therefore be converted to percentages and interpreted as percentage point changes attributable to PBF.

Indicator 1: Impact of PBF on the proportion of staff having been evaluated for their performance in the last year

Indicator measurement and calculation. The sample for this indicator was limited to clinical skilled health staff who had already worked at their current facility for a minimum of one year. Health workers were asked to indicate whether they had had a meeting with their internal or external supervisor to discuss the attainment of their objectives as mentioned in last year's "fiche / grille d'évaluation", which is to be filled and evaluated jointly by each health worker and their direct supervisor on a yearly basis. The indicator was calculated as the proportion of health workers who had responded positively to this question.

Results. Results pertaining to indicator 1 are displayed in **Box 1**. Overall, the proportion of health workers reporting performance evaluation remained relatively stable around approximately 70%, with slightly lower evaluation rates in intervention districts. Comparing health workers in intervention facilities to health workers in control facilities (specification 1), no impact of PBF could be detected. The effect estimate is negative around 5 percentage points (pp), but far from statistical significance. Comparing the different intervention arms to

Box 1: Proportion of staff having been evaluated for their performance in last year

Specification 1 (quasi-experiment): Intervention (pooled) vs control

Specification 3 (experiment): T2, T3, and T4 vs T1

BL n=872
EL n=1,248

All	-0.052

All	-0.025	-0.050	0.013

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	-0.097	0.037	0.009	-0.080

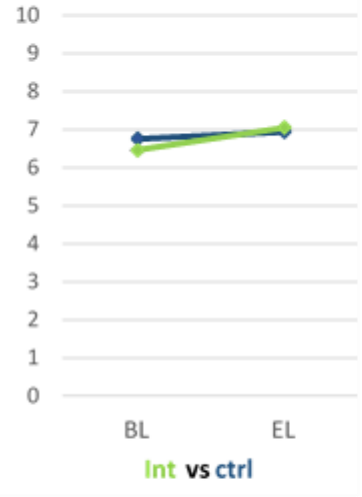
all controls (specification 2), no impact of T1, T2, or T3 was apparent either. Effect estimates for T2 and T3 are positive but close to zero, not reaching statistical significance. Estimates for T1 and T4 are somewhat larger and negative, but also not statistically significant. The latter is at least in part due to an above-average general decline in performance evaluation of around -14 pp in the Nouna district, where most of the T4 facilities are located, calling into question the suitability of using all control districts as counterfactual for this particular comparison. Comparing T2, T3, and T4 to T1 (specification 3) indicated no additional benefit of the various "add-ons" beyond the standard T1 in regards to the proportion of health workers who had been evaluated for their performance in the last year.

Indicator 2: Impact of PBF on health workers' perceived individual agency

Indicator measurement and calculation. Perceived individual agency refers to the extent to which health workers feel that they can influence what happens at their health facility. It was measured using two items, "I have significant influence on decisions affecting our facility." and "I have control over what happens at my facility." Respondents were asked to indicate their degree of agreement with the statements on a scale from 0 (complete disagreement) to

Box 2: Perceived individual agency

Specification 1 (quasi-experiment): Intervention (pooled) vs control



All	0.381

Specification 3 (experiment): T2, T3, and T4 vs T1

BL n=1,328
EL n=1,761

All	-0.077	0.126	0.198

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	0.336	0.318	0.520*	0.452

10 (complete agreement). Cronbach's alpha for the two items was acceptable at .65. The combined indicator was calculated as a respondent's unweighted of mean of responses to the two items.

Results. Results pertaining to indicator 2 are displayed in **Box 2**. Overall, perceived individual agency scores remained relatively stable between baseline and endline, between 6.5 and 7 on the 0-10 scale. Comparing health workers in intervention facilities to health workers in control facilities (specification 1), no impact of PBF could be detected, with a non-significant positive effect estimate of about 1/3 point on the 0-10 scale. Comparing the different intervention arms to all controls (specification 2), all effect estimates are positive and of similar magnitude, but only the one for T3 reached statistical significance (DID = 0.520). Effect estimates for the comparison of T2, T3, and T4 to T1 (specification 3) are close to zero and not statistically significant, indicating no additional benefit of the various "add-ons" beyond the standard T1 in regards to health workers' perceived individual agency.

Indicator 3: Impact of PBF on health workers' satisfaction with the physical work environment

Indicator measurement and calculation. Health workers' satisfaction with their physical work environment was assessed with seven items (satisfaction with availability of medication, availability of equipment, availability of consumables, availability of registers and forms, state of buildings, protection against risks (e.g. infection prevention), quality of services that can be delivered in light of the working conditions). Health workers were asked to indicate their degree of satisfaction with each of the seven aspects on a scale from 0 (not satisfied at all) to 10 (fully satisfied). Cronbach's alpha for the seven items was good at .81. The composite indicator was calculated as a respondent's unweighted mean of responses to the seven items.

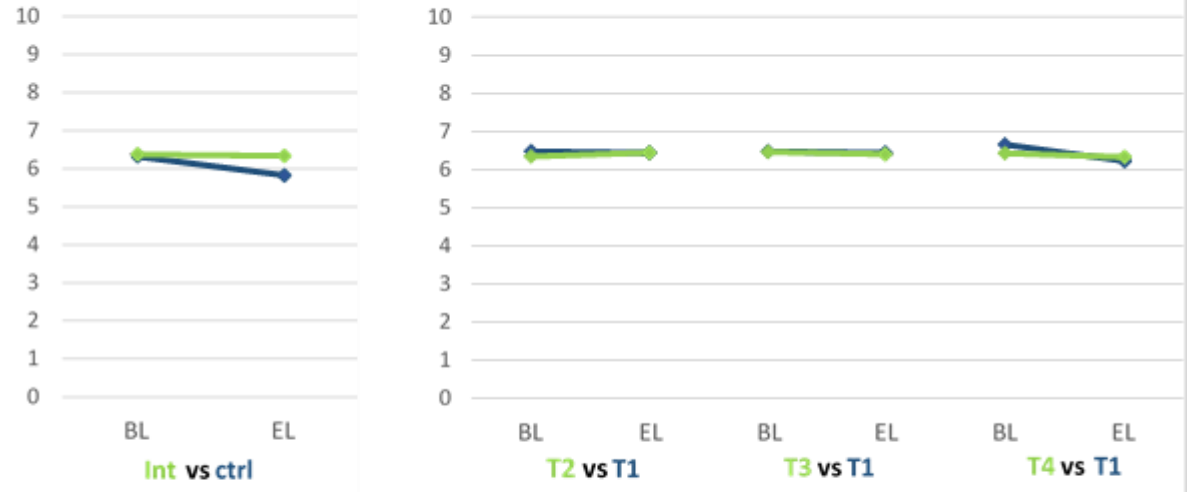
Results. Results pertaining to indicator 3 are displayed in **Box 3**. Overall, satisfaction with the physical work environment was moderate around 6 on the 0-10 scale. Satisfaction remained relatively stable between baseline and endline among intervention health workers, but decreased by around ½ point among health workers working in control facilities. Regression

Box 3: Satisfaction with the physical working environment

Specification 1 (quasi-experiment): Intervention (pooled) vs control

Specification 3 (experiment): T2, T3, and T4 vs T1

BL n=1,336
EL n=1,761



All	0.500*	All	0.179	0.028	0.287

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	0.450	0.646*	0.486	0.485*

results (specification 1) confirmed this difference to be statistically significant. Comparing the different intervention arms to all controls (specification 2), all effect estimates are positive and of similar magnitude, but only those for T2 and T4 reached statistical significance. Effect estimates for the comparison of T2, T3, and T4 to T1 (specification 3) are close to zero and not statistically significant, indicating no additional benefit of the various "add-ons" beyond the standard T1 in regards to health workers' satisfaction with their physical work environment.

Indicator 4: Impact of PBF on health workers' satisfaction with their compensation

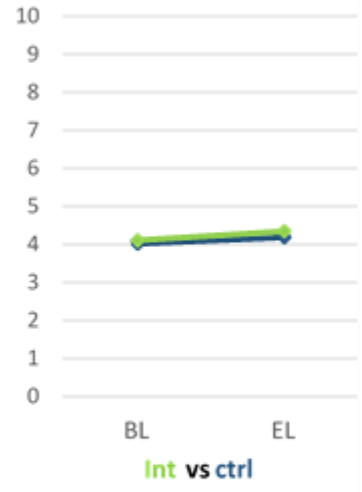
Indicator measurement and calculation. Health workers' satisfaction with their compensation was assessed with five items (satisfaction with chances of being financially or otherwise compensated for hard work, salary, diverse benefits, total revenue, housing). Health workers were asked to indicate their degree of satisfaction with each of the five aspects on a scale from 0 (not satisfied at all) to 10 (fully satisfied). Cronbach's alpha for the five items was good at .77. The composite indicator was calculated as a respondent's unweighted mean of responses to the five items.

Results. Results pertaining to indicator 4 are displayed in **Box 4**. Overall, satisfaction with compensation was moderate, between 4 and 4.5 on the 0-10 scale. Satisfaction remained fairly stable between baseline and endline for health workers from both intervention and control facilities. Comparing all intervention to all control health workers (specification 1), no impact of PBF is apparent; the effect estimate is very small and positive, but not statistically significant. Estimates for the comparison of the different intervention arms to all controls (specification 2) are also close to zero with the exception of a slightly larger estimate for T3, but all statistically insignificant. Effect estimates for the comparison of T2, T3, and T4 to T1 (specification 3) indicate no additional benefit of the various "add-ons" beyond the standard T1.

Additional analyses. We further performed all impact analysis separately on the two satisfaction items likely most directly influenced by PBF, namely "satisfaction with chances of being financially or otherwise compensated for hard work" and "satisfaction with total revenue". While none of the impact estimates for "satisfaction with total revenue" were statistically different from zero, results (not displayed in tabular form) indicate a small positive impact on "satisfaction with chances of being financially or otherwise compensated for hard work" in T3 and T4 compared to controls (specification 1). Restricting the analysis to only Boucle du Mouhoun for potentially more appropriate controls, the latter estimate reduces to only 0.284, however. In the experimental design part, only T3 appeared superior to T1.

Box 4: Satisfaction with compensation

Specification 1 (quasi-experiment): Intervention (pooled) vs control



All	0.143

Specification 3 (experiment): T2, T3, and T4 vs T1

BL n=1,326
EL n=1,760

All	0.063	0.342	-0.064

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	0.098	0.072	0.358	0.115

Indicator 5: Impact of PBF on health workers' satisfaction with management and supervision

Indicator measurement and calculation. Health workers' satisfaction with their management and supervision was assessed with six items (satisfaction with how the management team manages this health facility, opportunities to discuss issues pertaining to the health facility with the direct supervisor, opportunities to bring in new ideas, implication in decision-making at the COGES level, support received by the direct supervisor, recognition of one's work by the direct supervisor). Health workers were asked to indicate their degree of satisfaction with each of the six aspects on a scale from 0 (not satisfied at all) to 10 (fully satisfied). Cronbach's alpha for the six items was good at .87. The composite indicator was calculated as a respondent's unweighted mean of responses to the six items.

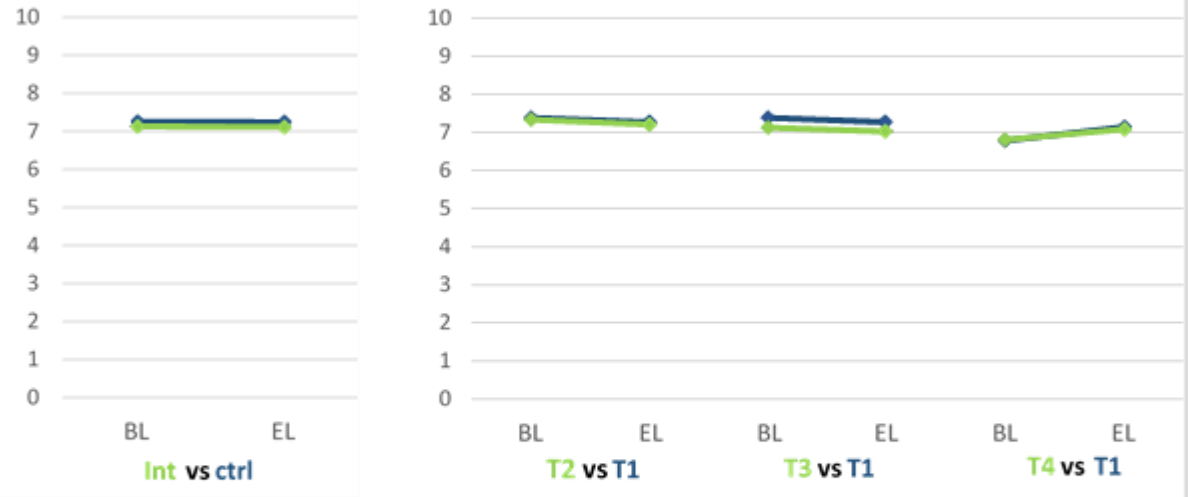
Results. Results pertaining to indicator 5 are displayed in **Box 5**. Overall, satisfaction with management and supervision was moderate, with scores between 7 and 7.5 on the 0-10 scale. Satisfaction remained almost perfectly stable between baseline and endline for health workers from both intervention and control facilities. The DID estimate for the comparison of all intervention health workers to all control health workers (specification 1) is close to zero and not statistically significant, accordingly. Estimates for the comparison of the different

Box 5: Satisfaction with the management of the health facility

Specification 1 (quasi-experiment): Intervention (pooled) vs control

Specification 3 (experiment): T2, T3, and T4 vs T1

BL n=1,331
EL n=1,760



All	0.018	**All**	0.036	0.050	-0.166

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	-0.004	-0.060	-0.048	0.330

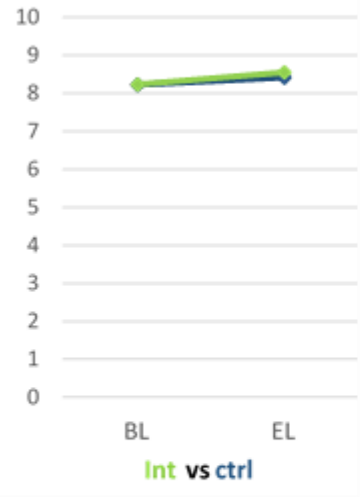
intervention arms to all controls (specification 2) are also close to zero and insignificant. This is with the exception of T4, for which the estimate is positive and a little larger, but also insignificant, and which is largely attributable to an above-average overall improvement in satisfaction with management and supervision in the Solenzo district. The comparison of T2, T3, and T4 to T1 (specification 3) indicates no additional benefit of the various "add-ons" beyond the standard T1.

Indicator 6: Impact of PBF on health workers' intrinsic motivation

Indicator measurement and calculation. Health workers' intrinsic motivation was assessed with the six intrinsic motivation and integrated/identified regulation items from the Self-Determination Theory-based scale developed in the context of this impact evaluation **Error! R eference source not found.**. Specifically, respondents were asked to indicate the extent to which diverse potential reasons for being motivated to do one's job were important to them personally, on a scale from 0 (not important at all) to 10 (very important). Reasons pertaining to intrinsic motivation included for instance "Because I enjoy doing what I do at work every day." or "Because I want to make a difference in people's lives." Cronbach's alpha for the six

Box 6: Intrinsic motivation

Specification 1 (quasi-experiment): Intervention (pooled) vs control



All	0.125

Specification 3 (experiment): T2, T3, and T4 vs T1

BL n=1,347
EL n=1,761

All	0.202	0.253	0.338

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	0.011	0.197	0.239	0.297***

items was good at .78. The composite indicator was calculated as a respondent's unweighted mean of responses.

Results. Results pertaining to indicator 6 are displayed in **Box 6**. Overall, intrinsic motivation scores were high, between 8 and 8.5 on the 0-10 scale. Scores increased slightly between baseline and endline in both the intervention and control group. The impact estimate for the comparison of all intervention to all control respondents (specification 1) is positive, but small and statistically insignificant. Effect sizes for the comparison of the different intervention arms to all controls (specification 2) were positive but small and insignificant for T1 and T2, and slightly larger for T3 and T4, with only T4 reaching statistical significance. Comparisons of T2, T3, and T4 with T1 (specification 3) yielded small positive but non-significant impact estimates, indicating no additional value of the various "add-ons" beyond the standard T1 in regards to health workers' intrinsic motivation.

Summary: Impact of PBF on human resource factors

Table 9 summarizes impact estimates for the six indicators pertaining to human resource factors. Positive and statistically significant impact estimates are marked in green, negative and significant impact estimates in red. Cells not marked in color contain estimates that did not reach statistical significance.

Table 9: Summary of results pertaining to the impact of PBF on human resources factors

	Quasi-experiment					Experiment		
	PBF vs control	T1 vs control	T2 vs control	T3 vs control	T4 vs control	T2 vs T1	T3 vs T1	T4 vs T1
1: Performance evaluation	-0.052	-0.097	0.037	0.009	-0.080	-0.025	-0.050	0.013
2: Perceived individual agency	0.381	0.336	0.318	0.520*	0.452	-0.077	0.126	0.198
3: Sat. with phys. work environm.	0.500*	0.450	0.646*	0.486	0.485*	0.179	0.028	0.287
4: Sat. with compensation	0.143	0.098	0.072	0.358	0.115	0.063	0.342	-0.064
5: Sat. with superv., managem.	0.018	-0.004	-0.060	-0.048	0.330	0.036	0.050	-0.166
6: Intrinsic motivation	0.125	0.011	0.197	0.239	0.297***	0.202	0.253	0.338

Note: * p < 0.1, ** p < 0.05, *** p < 0.01. All effect estimates pertain to absolute (as opposed to relative) change. Effect estimates for indicator 1 can be converted to percentages and reflect percentage point changes, whereas estimates for all indicators pertain to point changes on the 0-10 response scale.

3.4. Impact of PBF on the health service quality

In this section, result pertaining to the impact of PBF on the quality of health services are presented. Our understanding of quality of care follows Donabedian's [24] model to include both input and process elements. Specific indicators include:

7. Proportion of facilities with permanent availability of power and safe water in the last 7 days
8. Proportion of facilities with at least one unit of 23 essential drugs in stock
9. Proportion of observed ANC cases having received three key routine ANC services
10. Proportion of observed ANC cases having received patient education on three key elements
11. Proportion of children observed in curative consultations having been assessed for all IMCI danger signs
12. Proportion of children observed in curative consultations having been assessed for common childhood illness symptoms according to IMCI

13a. Proportion of ANC clients perceiving adequate quality of care on seven key elements

13b. Proportion of U5 consultation clients perceiving adequate quality of care on seven key elements
13c. Proportion of curative consultation clients aged 5 or older perceiving adequate quality of care on seven key elements

Data for indicators 7 and 8 were extracted from the health facility assessment; for indicators 9 and 10 from direct observations of ANC consultations; for indicators 11 and 12 from direct observations of consultations of children under the age of 5; and for indicators 13a-13c from exit interviews for the respective services.

All indicators in this section are binary, meaning that each facility, observed case, or client either was of/perceived adequate quality, or not. The effect estimates can therefore be converted to percentages and interpreted as percentage point changes attributable to PBF compared to status quo (specifications 1 and 2), or of intervention arms T2-T4 over and above T1 (specification 3).

Indicator 7: Impact of PBF on the proportion of facilities with permanent availability of power and safe water in the last 7 days

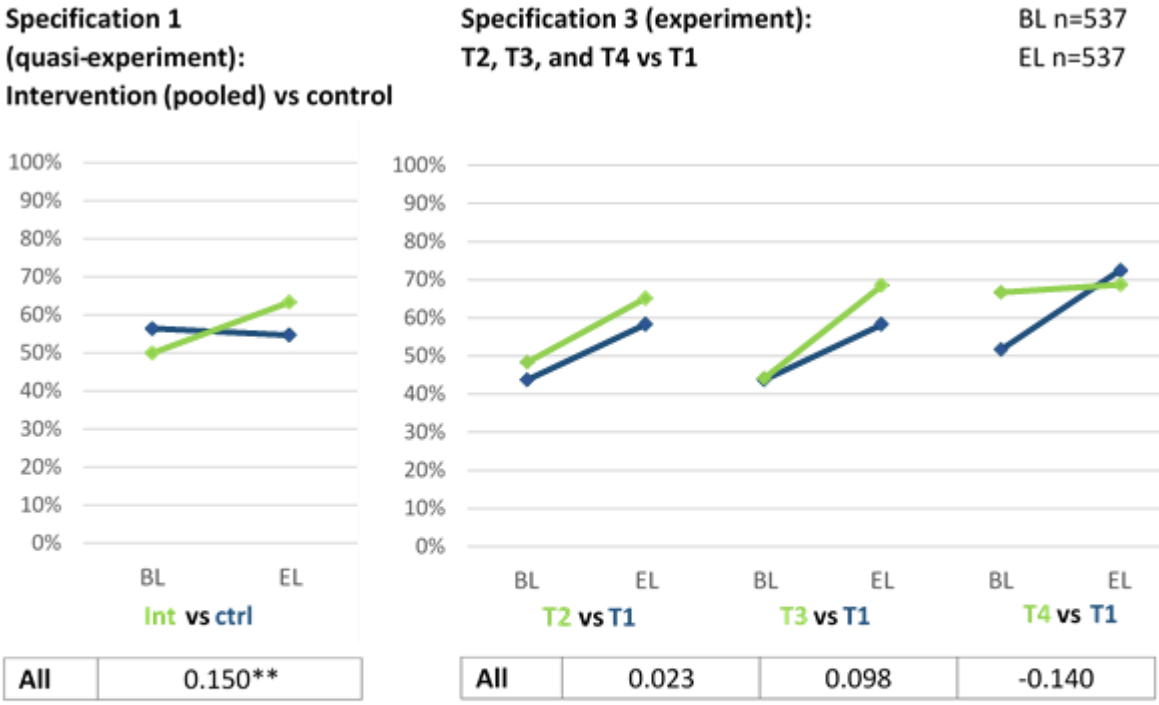
Indicator measurement and calculation. Data for this indicator were collected using the health facility assessment. In regards to electricity, all sources including solar power and generators in addition to power supply by the national electricity company were counted. Safe water was defined as "improved water source" as per the WHO definition[8]. A facility was counted as having had permanent availability only if they disposed of both electricity and safe water source and if both had been permanently supplied in the seven days prior to the survey.

Results. Results pertaining to indicator 7 are displayed in **Box 7**. Overall, availability of power and safe water remained stable around 55% between baseline and endline in the control facilities, while it increased from 50% to 63% in intervention facilities. The corresponding impact estimate (specification 1 and 2) was statistically significant, indicating 15 pp higher availability levels of power and safe water attributable to PBF overall as well as in all intervention arms but T4, particularly strongly in T3 (+26 pp). Comparisons of T2, T3 and T4 against the basic T1 in the experimental study component (specification 3), however, indicate no additional benefit of the various "add-ons" beyond the standard T1 in regards to the availability of power and safe water.

Additional analyses. Additional analyses run separately on the two components (i.e. electricity and water supply) indicate that the observed impacts were largely due to impact on availability of power, although some improvements on the availability of water were also apparent, in part due to decreased availability in control facilities.

[8] http://www.who.int/water_sanitation_health/monitoring/water.pdf

Box 7: Proportion of facilities with permanent availability of power and safe water in the last 7 days



All	0.150**

All	0.023	0.098	-0.140

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	0.114	0.186	0.261***	0.038

Indicator 8: Impact of PBF on the proportion of facilities with at least one unit of 23 essential drugs in stock

Indicator measurement and calculation. Data for this indicator were collected using the health facility assessment. The impact evaluation team defined a list of 23 individual drugs and/or groups of interchangeable drugs to capture availability of essential drugs across seven categories necessary for the delivery of basic health care services. The list of the drugs selected to compile this indicator was based on national and international guidelines and included:

1. Vaccines: BCG, OPV, Pentavalent, Measles
2. Antibiotics: Ciprofloxacine, Cotrimoxazole, Metronidazole, Amoxicilline
3. Antimalarials: Quinine (tablet), ACT adults (Artesunate-amodiaquine or Artemether-lumefantrine), ACT children (Artesunate-amodiaquine (tablet) or Artemether-lumefantrine (sirup))
4. Acute child care: ORS, Paracetamol, Quinine (injectable), Ampicillin, Ceftriaxone, Diazepam (injectable)

5. Antenatal care: Iron supplements, Sulfadoxine-pyrimethamine, Tetanus toxoid
6. Labor and delivery: Oxytocin
7. Family planning: Short-term methods (combination pill or progesterone only pill or DepoProvera), Long-term methods (Norplant or IUD)

A facility was only considered to have all essential drugs in stock if of each of the above 23 drugs or groups of drugs at least one unit was in stock on the day of the study team visit.

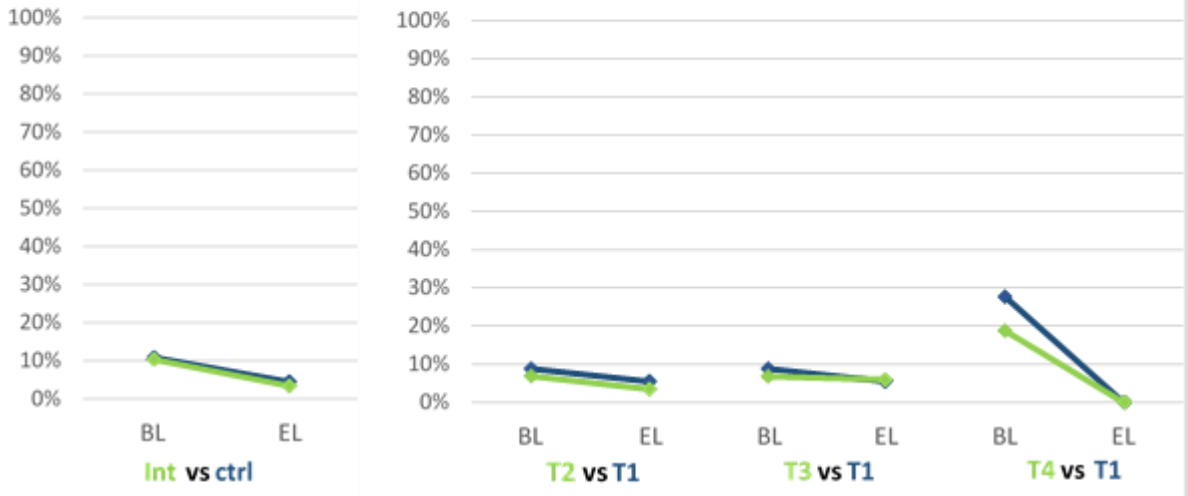
Results. Results pertaining to indicator 8 are displayed in **Box 8**. Overall, full availability of essential drugs was very low, 10/11% at baseline and decreasing to 3/4% at endline. Compared to status quo, no intervention effect could be detected overall (specification 1) nor for the single intervention arms T1, T2, and T3 (specification 2). In contrast, comparing T4 to controls suggests a significant negative impact of PBF (-12.5 pp). However, this can be largely attributable to a comparatively sharp drop in drug availability levels in the Nouna and Solenzo districts overall (-18 pp and -27 pp, respectively) as visible from the T4 vs T1 graph in **Box 8**, calling into question the appropriateness of using all control districts as counterfactual for this particular comparison. Comparing T2, T3, and T4 to T1 in the experimental study component (specification 3), no additional benefit of the various "add-ons" beyond the standard T1 is apparent in regards to the availability of essential drugs.

Box 8: Proportion of facilities with at least one unit of all essential drugs in stock

Specification 1 (quasi-experiment): Intervention (pooled) vs control

Specification 3 (experiment): T2, T3, and T4 vs T1

BL n=476
EL n=525



All	-0.016

All	0.050	0.063	-0.017

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	-0.034	0.035	0.049	-0.125***

Additional analyses. Analyses using a drug availability score instead of the "all-or-nothing" indicator (see 2.6) led to very similar results regarding the impact of PBF, but underlined that although full availability of all essential drugs was rare, facilities did have most essential drugs in stock, on average about 19 out of 23 at both baseline and endline. Drugs with particularly low availability rates were SP (65% of facilities out of stock at baseline, 45% at endline), iron supplements (45% of facilities out of stock at endline), paracetamol sirup (50% of facilities out of stock at endline), and ACT drugs for children (30% of facilities out of stock at baseline, 58% at endline).

Indicator 9: Impact of PBF on the proportion of observed ANC cases having received three key routine ANC services

Indicator measurement and calculation. Data for this indicator were collected using direct observations of antenatal care consultations. The three elements below were selected as key indications of ANC quality of service delivery, since they are to be provided to every woman at every ANC visit (whereas the provision of other services depends on the timing of the visit in the course of the pregnancy as well as on potential prior visits):

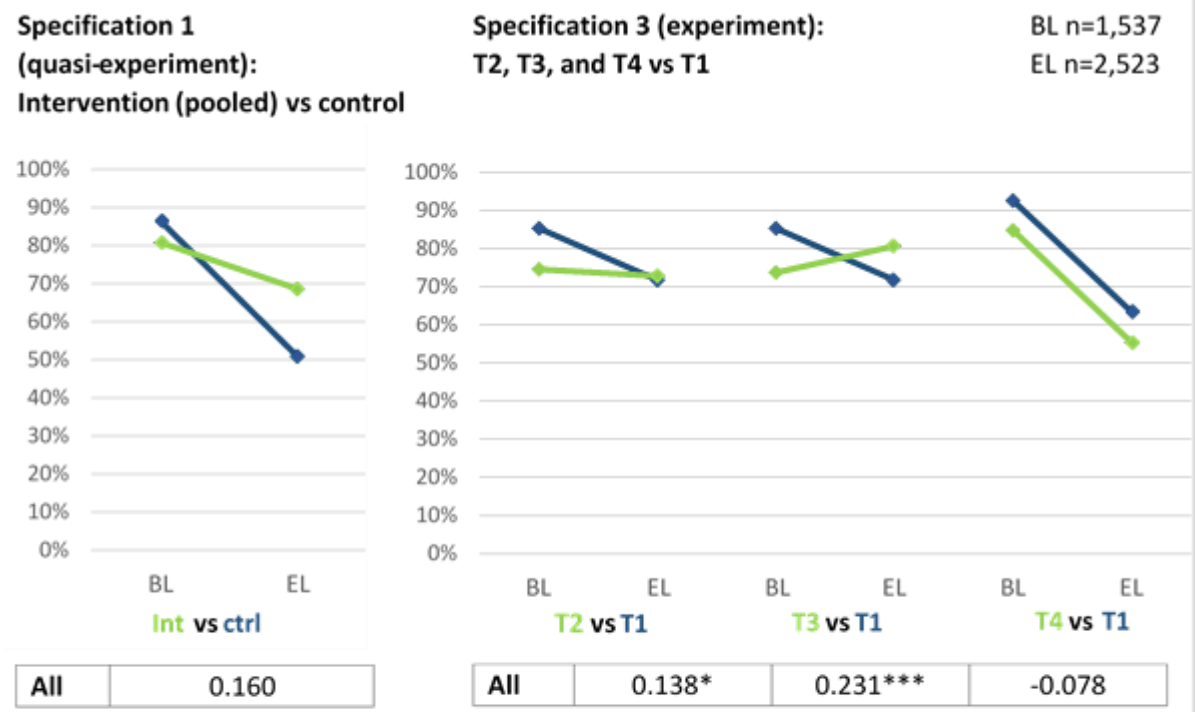
- Weight measurement
- Blood pressure measurement
- Prescription of iron supplements

The indicator was calculated as "all-or-nothing", i.e. only observed consultations in which all three services were provided were considered as of good quality.

Results. Results pertaining to indicator 9 are displayed in **Box 9**. Overall, service quality was relatively high at baseline, with all three key services provided in 82% of all observed cases. At endline, the proportion of observed cases of adequate quality dropped to 69% in intervention facilities and 51% control facilities. The corresponding estimate for the effect of PBF compared to control (specification 1) is positive at around 16 pp, but does not reach statistical significance due to substantial variation between districts. The overall positive effect estimate is primarily driven by intervention arms T2 and T3 (specification 2), for which a positive impact compared to status quo of could be detected (T2: +27 pp; T3: +33 pp). The comparison of T2, T3, and T4 against the basic T1 (specification 3) supports this finding. Whereas quality decreased over time in T1 facilities, it remained fairly stable in T2 and increased in T3. Results therefore indicate an additional benefit of 14 pp for T2 and of 23 pp for T3 compared to T1. No additional benefit of T4 could be detected.

Additional analyses. Additional separate analyses for each subcomponent show that the decrease in quality at endline was primarily due to a decrease in iron supplement prescription from 91% at baseline to 72% and 53% at endline in intervention and control cases, respectively. In contrast, weight and blood pressure measurement were done in almost all cases both at baseline and endline.

Box 9: Proportion of observed ANC cases having received all three key routine ANC services



All	0.160

All	0.138*	0.231***	-0.078

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	0.081	0.267*	0.333**	-0.086

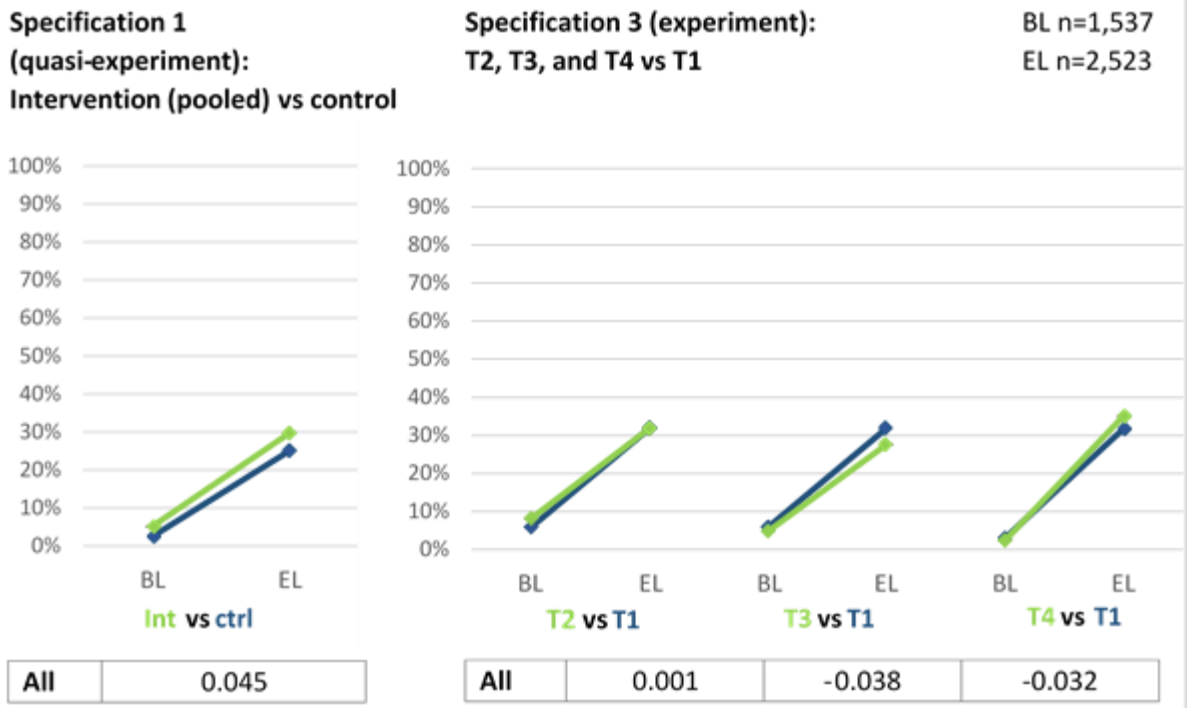
Indicator 10: Impact of PBF on the proportion of observed ANC cases having received patient education on three key elements

Indicator measurement and calculation. Data for this indicator were collected using direct observations of antenatal care consultations. Similar to indicator 9, the following patient education aspects were selected as key elements on the basis of the fact that they should be provided to every woman upon every ANC visit:

- At least two out of five pregnancy danger signs (vaginal bleeding, fever, fatigue or excessive shortness of breath, hand and face swelling, intense headache or impaired eyesight)
- Nutrition in pregnancy and breastfeeding
- Importance of skilled birth attendance and/or having a birth plan in place.

The indicator was calculated as “all-or-nothing”, i.e. only observed consultations in which patient education on all three services were provided were considered as adequate.

Box 10: Proportion of observed ANC cases having received patient education on all three key elements



All	0.045

All	0.001	-0.038	-0.032

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	0.045	0.048	0.017	0.113*

Results. Results pertaining to indicator 10 are displayed in **Box 10**. Overall, service quality in regards to patient education was extremely low at baseline, with all elements being provided only in 5% of the observed cases. Quality increased substantially at endline, with 29% of all observed consultations including adequate patient education, relatively evenly so across the intervention and control groups. Accordingly, comparing PBF to status quo (specification 1 and 2), no intervention effect could be detected overall nor for intervention arms T1, T2, and T3 specifically. Results of the comparison of T2 and T3 to the standard T1 (specification 3) indicate no additional benefit of the T2 and T3 "add-ons" in regards to patient education. In contrast, they suggest a positive effect of intervention arm T4 compared to status quo (specification 2; +11 pp). However, as the comparison of T4 to T1 in the experimental part of the design (specification 3) shows, this result is a reflection of the above-average overall positive development in the Nouna and Solenzo districts where all T4 facilities are located, rather than better performance of T4 compared to T1 in regards to quality improvements.

Additional analyses. Additional separate analyses for each subcomponent included in the patient education indicator show that very low quality at baseline was largely due to nutrition counselling being done in only 10% of observed cases, whereas rates were between 30% and 40% for the other elements (pregnancy danger signs and skilled birth attendance). At endline,

rates were around 50% or higher for all three elements. Results also show that the overall lack of effects masks a negative impact of PBF specifically on pregnancy danger signs overall as well as for intervention arms T1, T2, and T3 compared to status quo. This negative impact is attributable to a substantially larger increase in danger sign counseling among cases observed in control facilities (BL: 29%, EL: 74%) compared to cases observed in intervention facilities (BL: 41%, EL: 56%).

Indicator 11: Impact of PBF on the proportion of children observed in curative consultations having been assessed for all IMCI danger signs

Indicator measurement and calculation. Data for this indicator were collected using direct observations of curative consultations of children under the age of 5. In line with the IMCI guidelines[9], observed consultations were considered to be of good quality if all of the four following assessments were done:

- Provider asks whether child is able to drink or breastfeed
- Provider asks whether child vomits everything they consume
- Provider asks for symptoms of lethargy or change in level of consciousness
- Provider asks for symptoms of convulsion

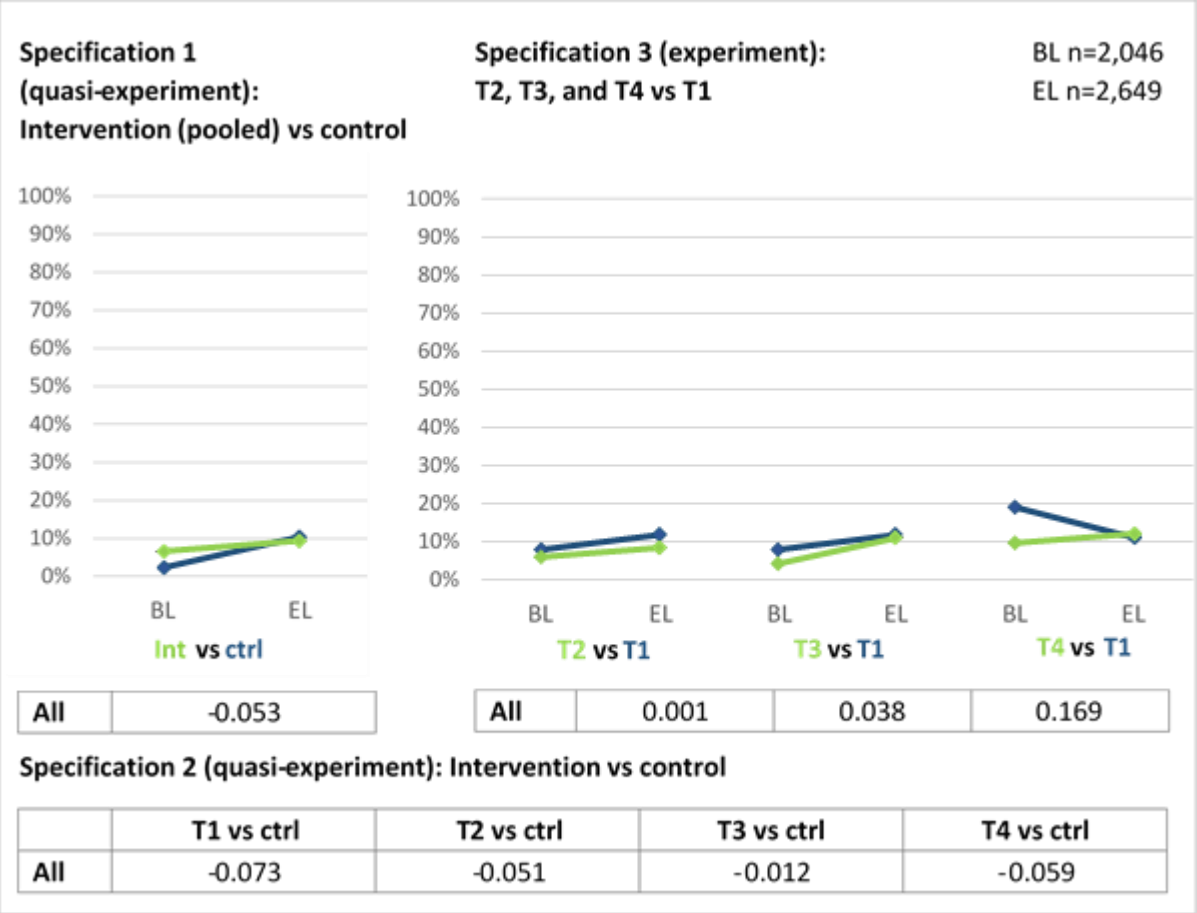
The indicator was calculated as "all-or-nothing", i.e. only observed consultations in which all four assessments took place were considered of good quality.

Results. Results pertaining to indicator 11 are displayed in **Box 11**. IMCI danger signs assessment rates were very low at baseline (6%) and only slightly higher at endline (10%). Rates were largely similar between the intervention and control group as well as between the different intervention arms, so that no impact could be detected, neither for PBF compared to status quo, nor within the experimental study component.

Additional analyses. Additional analyses show that assessment for lethargy and convulsion was done infrequently at both data collection time points (assessment rates between 5% and 30%), whereas assessment of drinking and vomiting was more frequent (rates around 50%). Separate impact analyses on the four sub-components included in the indicator show statistically significant negative impact of PBF compared to status quo overall as well as for intervention arms T1 and T2 in regards to assessment for vomiting and for symptoms of lethargy.

[9] http://www.who.int/maternal_child_adolescent/documents/IMCI_chartbooklet/en/

Box 11: Proportion of children observed in curative consultations having been assessed for all IMCI danger signs



All	-0.053

All	0.001	0.038	0.169

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	-0.073	-0.051	-0.012	-0.059

Indicator 12: Impact of PBF on the proportion of children observed in curative consultations having been assessed for symptoms of common childhood illnesses according to IMCI

Indicator measurement and calculation. Data for this indicator were collected using direct observations of curative consultations of children under the age of 5. In line with the IMCI guidelines[10], observed consultations were considered to be of good quality if the provider assessed the child for the following:

- Provider asks for presence of fever
- Provider asks for presence of cough
- Provider asks for presence of diarrhea
- Provider asks for presence of ear problems
- Provider checks weight
- Provider checks temperature
- Provider checks for signs of anemia (conjunctivae or palms)
- Provider checks vaccination status

[10] http://www.who.int/maternal_child_adolescent/documents/IMCI_chartbooklet/en/

The indicator was calculated as “all-or-nothing”, i.e. only observed consultations in which all eight assessments took place were considered as of good quality.

Results. Results pertaining to indicator 12 are displayed in **Box 12**. Assessment rates were very low at baseline (5%) and increased to 16% in the intervention group and 34% in the control group at endline. This differential increase is reflected in a statistically significant estimate for the impact of PBF against status quo of around -23 pp (specification 1), which appears relatively uniform across the four intervention arms (specification 2). No additional benefit of the various “add-ons” in T2, T3, and T4 compared to the basic T1 is apparent from the results (specification 3). Important to note is that the comparatively larger increase in the control districts was driven specifically by a sharp increase in two districts, Barsalogho (+96 pp) and Ziniaré (+91%). If these two districts are excluded from the analyses, impact estimates remain negative, but decrease to around -5 pp and become statistically insignificant.

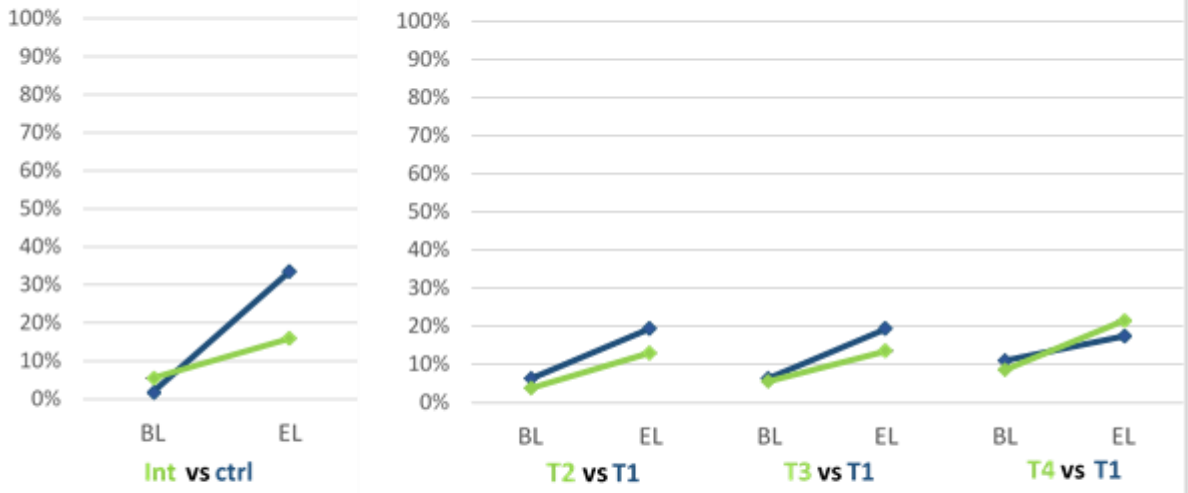
Additional analyses. The low assessment rates on the combined indicator are in particular due to low assessment rates for ear problems (23% at baseline, 44% at endline) and for vaccination status assessment (43% at baseline, 54% at endline). Assessment rates were moderate at around 65% at baseline for diarrhea, weight, and anemia. For diarrhea and

Box 12: Proportion of children observed in curative consultations having been assessed for symptoms of common childhood illnesses according to IMCI

Specification 1 (quasi-experiment): Intervention (pooled) vs control

Specification 3 (experiment): T2, T3, and T4 vs T1

BL n=2,046
EL n=2,649



All	-0.232*

All	-0.020	-0.018	0.009

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	-0.228*	-0.252*	-0.243*	-0.180

weight, they increased to around 87% at endline, whereas they remained at 68% for anemia. For all other elements, rates were above 70% at both baseline and endline. Separate impact analyses on the different sub-indicators show significant negative impact only for assessment of anemia, in particular in intervention arm T1 compared to controls.

Indicator 13a: Impact of PBF on the proportion of ANC clients perceiving adequate quality of care on seven key elements

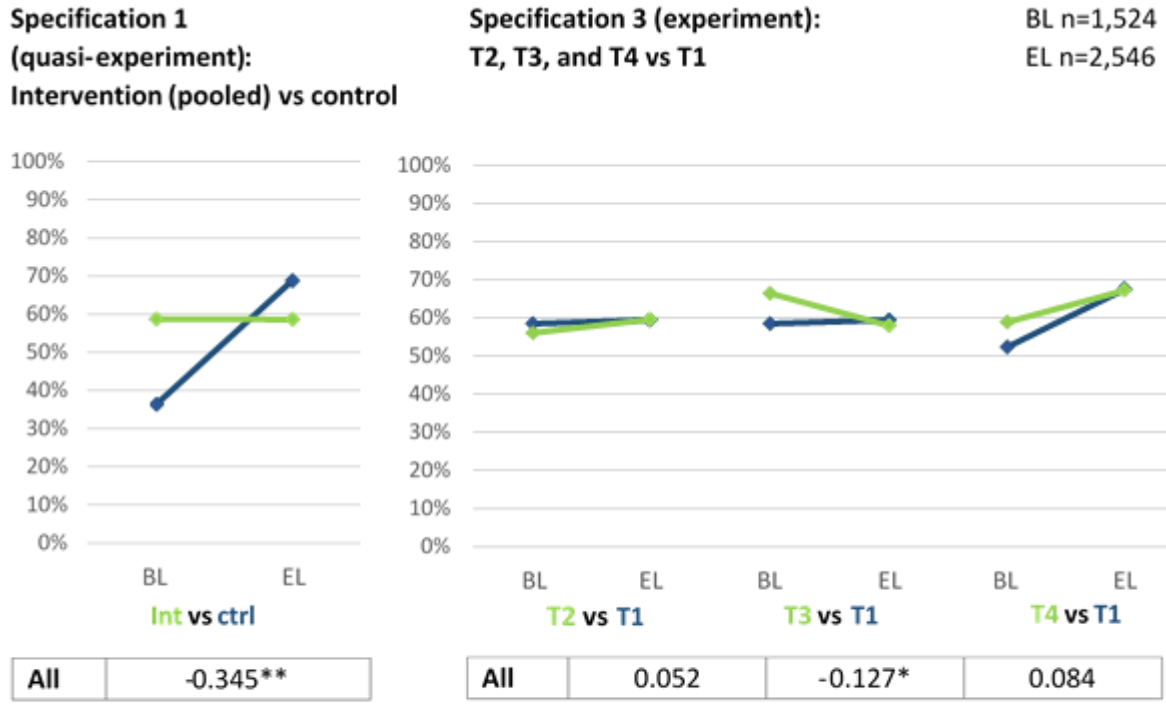
Indicator measurement and calculation. Data for this indicator were collected using exit interviews with antenatal care clients. Clients were asked to indicate the extent to which they were satisfied with a variety of aspects related to the care they had just received. From this list, we selected the seven aspects that applied to all three services (and pertained to all patients) for which exit interviews were conducted (ANC, consultations for children under 5 (indicator 13b), consultations for patients aged 5 and above (indicator 13c)) to allow comparison across all services. For instance, we did not include "prescribed medication was easy to obtain" or "consultation fees were reasonable" as these items would have been applicable only to the small portion of the sample who was prescribed drugs or payed consultation fees. Specifically, the following items were included:

- The health facility is clean.
- The health facility staff is polite and respectful.
- The health workers have explained your health status well.
- The time you spent waiting to be seen by a health worker was reasonable.
- You had sufficient privacy during your consultation/visit.
- The health worker spent enough time with you.
- The health facility opening hours correspond to your needs.

For each item, respondents were asked to indicate whether they agreed, neither agreed nor disagreed, or disagreed. We calculated the composite indicator as "all-or-nothing", meaning that only respondents who had agreed with all seven statements were treated as having perceived quality of care to be of adequate standards. In addition, as a robustness check, to the "all-or-nothing" indicator, we also calculated a score as the number of items to which the client had agreed, ranging from 0 (no agreement whatsoever) to 7 (agreement with all).

Results. Results pertaining to indicator 13a are displayed in **Box 13a**. Results of the comparison of PBF to status quo (specification 1) imply a strong negative intervention effect on perceived quality of ANC care at approximately -35 pp overall. It must be noted, however, that this negative effect was entirely driven by an increase in the proportion of clients who perceived adequate quality of care in the control group, with the related indicator moving from 36% at baseline to 69% at endline, whereas the proportion remained stable over time in the intervention group, at 59% overall. Wild bootstrapping further showed that the estimate is not robust and should be interpreted with extremely caution. Comparing the different intervention arms (specifications 2 and 3), there appears to have been little change in intervention arms T1 and T2, a small downward trend in T3 both compared to controls and to T1, and a small upward trend in T4 facilities. The latter reflects the general upward trend

Box 13a: Proportion of ANC clients perceiving adequate quality of care on seven key elements



All	-0.345**

All	0.052	-0.127*	0.084

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	-0.353**	-0.297*	-0.453***	-0.174

in the Nouna and Solenzo districts, however. No significant effect of T4 over and above the standard T1 could be detected.

Additional analyses. It is important to note is that while many respondents did not report adequate quality of care on all seven items (across time points and across intervention and control arms), most respondents perceived most single aspects to be of adequate quality. In the intervention group, respondents on average agreed to slightly more than six out of the seven statements at baseline and endline. In the control group, the average increased from around 5 to around 6.5 between baseline and endline. DID analyses using the score led to similar results than those using the "all-or-nothing" indicator. Interestingly, no single item was responsible neither for the shortfall from perfect perceived quality of care, nor for the increase in the control group. Rather, "non-agreements" were relatively evenly distributed across items for those respondents not perceiving full quality of care.

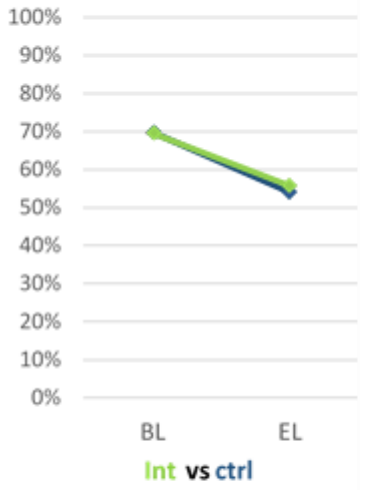
Indicator 13b: Impact of PBF on the proportion of U5 consultation clients perceiving adequate quality of care on seven key elements

Indicator measurement and calculation. Data for this indicator were collected using exit interviews with caregivers of curative care clients under the age of five. The indicator was measured and calculated analogous to indicators 13a.

Results. Results pertaining to indicator 13b are displayed in **Box 13b**. Overall, perceived quality of care of U5 consultations declined from 70% of caregivers perceiving adequate care at baseline to 55% at endline in both the intervention and control groups. The corresponding DID analyses (specifications 1 and 2) confirm no impact of PBF compared to status quo, neither overall, nor for the different intervention arms. The significant positive estimate for T4 compared to status quo in specification 2 is due a general increase in perceived quality of care in the Nouna and Solenzo districts as visible from the T4 vs T1 graph, calling into question the appropriateness of the counterfactual for this particular comparison. In the experimental study component, no additional benefit of T2, T3, and T4 over and above the basic T1 could be detected (specification 3).

Box 13b: Proportion of U5 consultation clients perceiving adequate quality of care on seven key elements

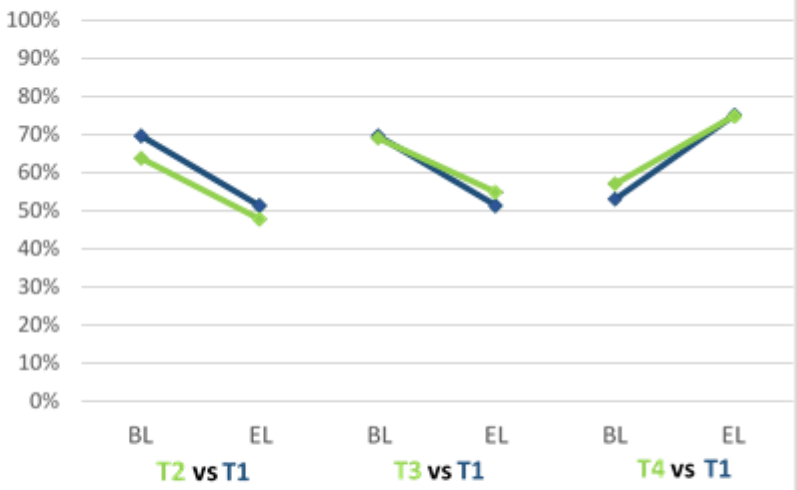
Specification 1 (quasi-experiment): Intervention (pooled) vs control



All	-0.010

Specification 3 (experiment): T2, T3, and T4 vs T1

BL n=2,023
EL n=2,617



All	0.007	-0.003	0.034

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	-0.063	-0.026	-0.029	0.305***

Additional analyses. Analyses using the score instead of the "all-or-nothing" indicator confirm these findings regarding the impact of PBF. Overall, most caregivers perceived high quality of care on most of the items. The average number of items to which respondents agreed declined only slightly between baseline and endline from 6.41 to 6.27 overall, reflecting the drop in the proportion of caregivers with full perceived quality of care visible in the "all-or-nothing" indicator. As for indicator 13a, no individual item was responsible for the decline in perceived quality of care.

Indicator 13c: Impact of PBF on the proportion of curative consultation clients aged 5 or older perceiving adequate quality of care on seven key elements

Indicator measurement and calculation. Data for this indicator were collected using exit interviews with curative care clients aged 5 or older, or their respective caregivers. The indicator was measured and calculated analogous to indicators 13a.

Results. Results pertaining to indicator 13c are displayed in **Box 13c**. Overall, perceived quality of care of curative consultations for children over the age of 5 and adults declined from 66% of clients/caregivers perceiving adequate care at baseline in both groups, to 57% at endline in the intervention group while remaining stable in the control group. The corresponding DID estimates (specification 1 and 2) are not statistically significant, however, neither overall, nor for any of the different intervention arms, although relatively large especially for T1 at -17 pp. The experimental study component shows no added benefit of T2 and T3 compared to the basic T1 (specification 3). However, results indicate a negative effect of T4 over and above T1. Again, this needs to be interpreted in light of a generally positive trend in the Nouna and Solenzo districts, more so in T1 than in T4, resulting in the negative effect estimate.

Additional analyses. Analyses using a perceived quality score instead of the "all-or-nothing" indicator confirm these findings regarding the impact of PBF. Overall, most caregivers perceived high quality of care on most of the items. The average number of items to which respondents agreed was 6.3 at baseline and endline. As for indicators 13a and 13b, "non-agreements" were relatively evenly distributed across items for those respondents not perceiving full quality of care.

Summary: Impact of PBF on health service quality

Table 10 summarizes impact estimates for the nine indicators pertaining to health service quality. Positive and statistically significant impact estimates are marked in green, negative and significant impact estimates in red. Cells not marked in color contain estimates that did not reach statistical significance.

Box 13c: Proportion of curative consultation clients aged 5 or older perceiving adequate quality of care on seven key elements

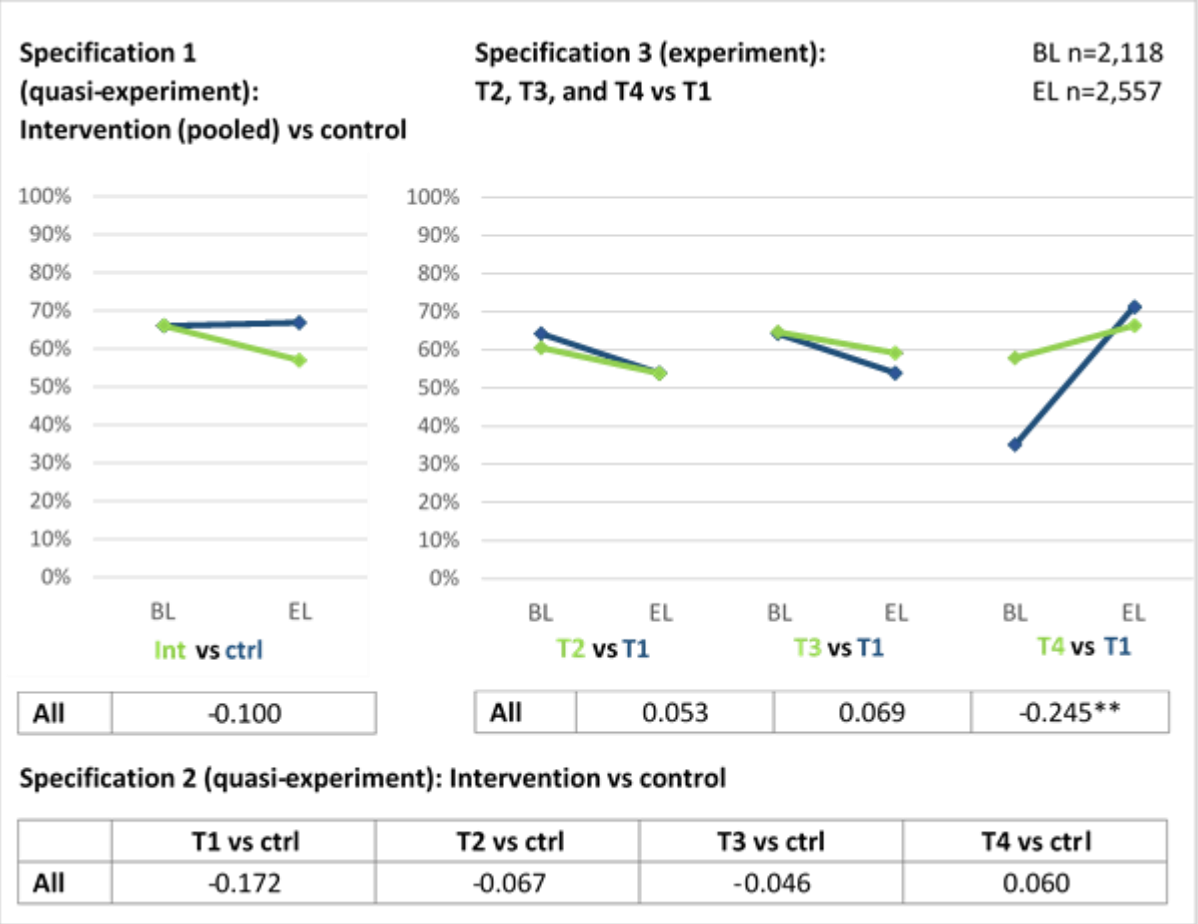


Table 10: Summary of results pertaining to the impact of PBF on health service quality

	Quasi-experiment					Experiment		
	PBF vs control	T1 vs control	T2 vs control	T3 vs control	T4 vs control	T2 vs T1	T3 vs T1	T4 vs T1
7: Availability of power & water	0.150**	0.114	0.186	0.261***	0.038	0.023	0.098	-0.140
8: Availability of essential drugs	-0.016	-0.034	0.035	0.049	-0.125***	0.050	0.063	-0.017
9: ANC routine services	0.160	0.081	0.267*	0.333**	-0.086	0.138*	0.231***	-0.078
10: ANC patient education	0.045	0.045	0.048	0.017	0.113*	0.001	-0.038	-0.032
11: IMCI danger signs	-0.053	-0.073	-0.051	-0.012	-0.059	0.001	0.038	0.169
12: IMCI routine symptoms	-0.232*	-0.228*	-0.252*	-0.243*	-0.180	-0.020	-0.018	0.009
13a: Perceived quality ANC	-0.345**+	-0.353**	-0.297*	-0.453***	-0.174	0.052	-0.127*	0.084
13b: Perceived quality U5	-0.010	-0.063	-0.026	-0.029	0.305***	0.007	-0.003	0.034
13c: Perceived quality 5+	-0.100	-0.172	-0.067	-0.046	0.060	0.053	0.069	-0.245**

Note: * p < 0.1, ** p < 0.05, *** p < 0.01. All effect estimates pertain to absolute (as opposed to relative) change. Effect estimates can be converted to percentages and reflect percentage point changes. + estimate is not robust according to wild bootstrap

3.5. Impact of PBF on the utilization of reproductive health care services

In this section, result pertaining to the impact of PBF on the utilization of maternal and reproductive health care services are presented. Specific indicators included:

14. Proportion of recently pregnant women with at least four ANC visits
15. Proportion of recently pregnant women with an ANC visit within first four months of pregnancy
16. Proportion of recently pregnant women having received at least 2 doses of tetanus vaccine during pregnancy
17. Proportion of recently pregnant women having been offered HIV testing during pregnancy
18. Number of HIV-positive mothers who have completed prophylactic ARV treatment (SNIS)
19. Proportion of recently pregnant women who have delivered in a formal health facility
20. Proportion of recently pregnant women with at least one PNC visit within 6 weeks after delivery
21. Proportion of recently pregnant women with at least three PNC visits within 6 weeks after delivery
22. Proportion of non-pregnant women aged 15-49 who use modern family planning methods

Data for the calculation of all indicators but 18 were extracted from the women's module of the household survey, which was administered to all women aged 15-49 in sampled households. The module contained questions pertaining to pregnancy and birth history, to utilization of maternal health care services in case of a pregnancy in the last two years, and to family planning. For indicator 18, routine health facility data (SNIS) were used, due to the absence of relevant information in our primary data collection tools.

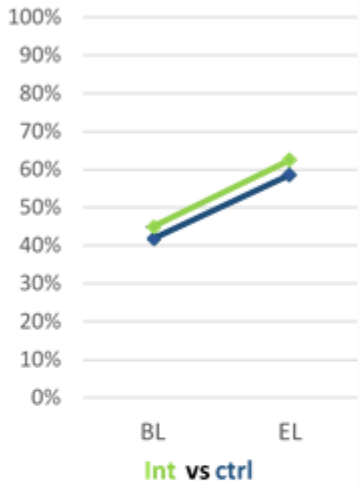
Indicator 14: Impact of PBF on the proportion of recently pregnant women with at least four ANC visits

Sample, indicator measurement and calculation. The sample was restricted to women who had ended a pregnancy within the 24 months prior to the interview date, irrespective of pregnancy outcome. The indicator was calculated as the proportion of the sample who reported to have completed four or more ANC visits at a formal care health facility.

Main results. Results pertaining to indicator 14 are displayed in **Box 14**. Overall, the proportion of women with at least four ANC visits increased from 44% at baseline to 62% at endline, with only small differences between the intervention and control groups. Accordingly, in the corresponding DID analysis (specification 1), no impact of PBF could be detected. Similarly, no effect of T1, T2, and T3 compared to controls (specification 2) could be detected; impact estimates are close to zero or slightly positive. T4 had a statistically significant negative impact. This is largely due to the fact that the increase in utilization of four ANC visits was not quite as steep as in the Nouna and Solenzo districts, where all T4

Box 14: Proportion of recently pregnant women with at least four ANC visits

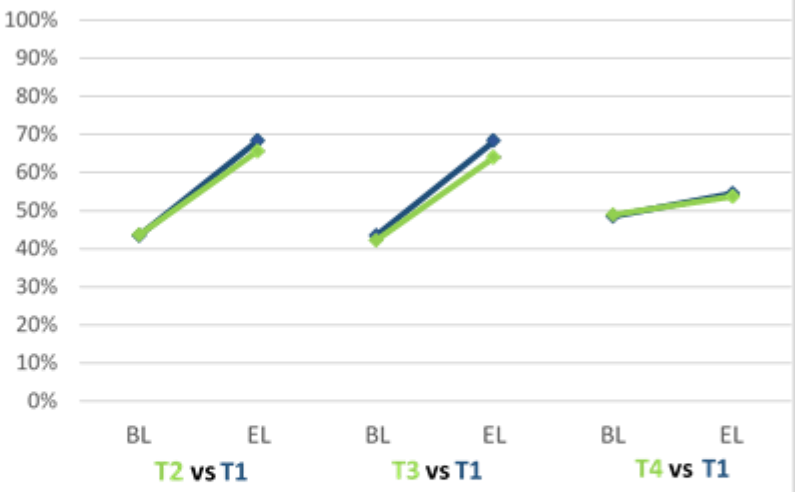
Specification 1 (quasi-experiment): Intervention (pooled) vs control



All	-0.004
Poor	0.037

Specification 3 (experiment): T2, T3, and T4 vs T1

BL n=6,565
EL n=6,365



All	-0.026	-0.018	-0.031
Poor	-0.038	-0.026	0.020

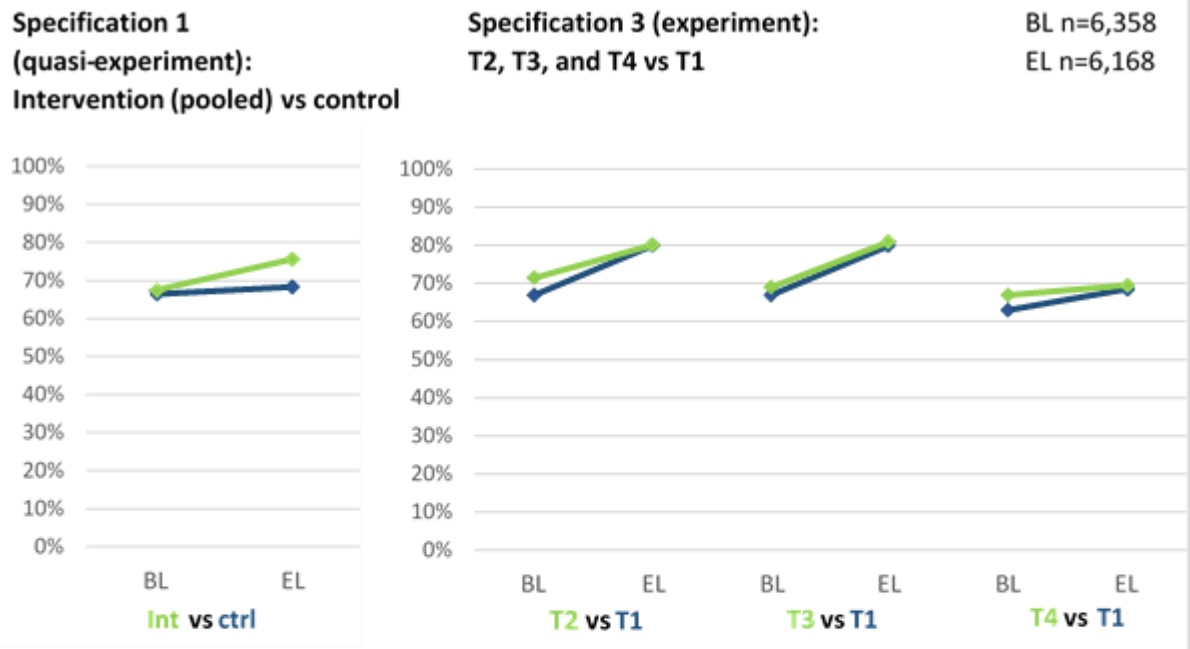
Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	-0.011	0.038	0.046	-0.132***
Poor	0.031	0.111	0.125**	-0.162**

facilities are located, as in the overall sample, calling into question the suitability of using all control districts as counterfactual for this particular comparison. The comparison of T2, T3, and T4 to T1 (specification 3) indicates no additional benefit of the "adds-on" compared to the standard T1.

Stratified analysis on the poorest 20%. Results using the subsample in the lowest asset index quintile largely confirm those of the overall analysis. Effect estimates are generally larger than for the overall sample, reaching statistical significance only for T3 compared to controls (DID = 0.125).

Box 15: Proportion of recently pregnant women with an ANC visit within first four months of pregnancy



All	0.057
Poor	0.031

All	-0.036	-0.007	-0.064
Poor	-0.071	-0.037	0.073

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	0.050	0.069	0.097	0.003
Poor	0.017	0.023	0.057	0.048

Indicator 15: Impact of PBF on the proportion of recently pregnant women with an ANC visit within first four months of pregnancy

Sample, indicator measurement and calculation. The sample was restricted to women who had ended a pregnancy within the 24 months prior to the interview date, irrespective of pregnancy outcome and who had completed at least one ANC visit in a formal health facility. The indicator was calculated as the proportion of the sample who reported to have had at least one ANC visit at a formal health facility within the first 4 months of pregnancy.

Main results. Results pertaining to indicator 15 are displayed in **Box 15**. Overall, the proportion of women with one or more of their ANC visits in their first trimester increased from 67% at baseline to 76% in the intervention group, while remaining almost stable in the control group. The corresponding DID estimate (specification 1) indicates an impact of +5.7 pp, which was not statistically significant from zero. In the comparison of the different

intervention arms against status quo (specification 2), coefficients for T2 and T3 are slightly higher, but also non-significant (+6.9 and +9.7 pp, respectively), while the coefficient for T4 is near zero. In the experimental study component (specification 3), however, no additional benefit of the "adds-on" could be confirmed.

Stratified analysis on the poorest 20%. Results using the subsample in the lowest asset index quintile confirm those of the overall analysis. Effect estimates are slightly lower than in the overall sample for the comparison of PBF to status quo overall and for T1, T2, and T3, as well as larger for T4, but all statistically insignificant.

Indicator 16: Impact of PBF on the proportion of recently pregnant women having received at least 2 doses of tetanus vaccine during pregnancy

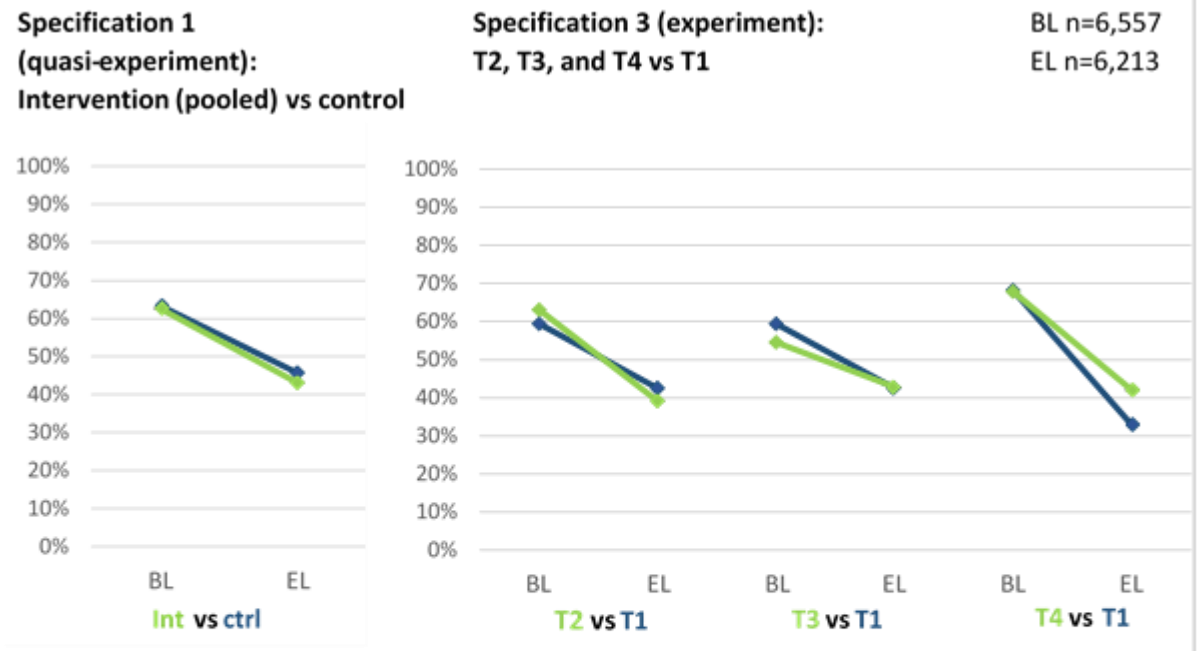
Sample, indicator measurement and calculation. The sample was restricted to women who had ended a pregnancy within the 24 months prior to the interview date, irrespective of pregnancy outcome. The indicator was calculated as the proportion of the sample who reported to have received at least two doses of tetanus vaccine during their pregnancy.

Main results. Results pertaining to indicator 16 are displayed in **Box 16**. Overall, the proportion of women who had received two doses of tetanus vaccine during pregnancy declined from 63% to 44% between baseline and endline[11], with only small differences between the intervention and control group overall. The corresponding DID estimate (specification 1) is near zero. Coefficients are somewhat larger for intervention arms T2, T3, and T4 compared to status quo, although all statistically insignificant (specification 2). Specifically, estimates indicate a positive effect of T3 (+7.5 pp) and negative effects of T2 and T4 (-7.5 and -8.5 pp, respectively). The latter is again influenced by a somewhat steeper decline in vaccination rates in the Nouna and Solenzo districts compared to the overall sample. Results for specification 2 are reflected in the experimental study component (specification 3), which shows a negative effect of T2 beyond the standard T1 (-7 pp) and a positive but statistically insignificant estimate for the comparison of T3 to T1. In contrast, the comparison of T4 to T1 in the Nouna and Solenzo districts indicates an added benefit of the option of community-based health insurance over and above the standard PBF in T1.

Stratified analysis on the poorest 20%. Results using the subsample in the lowest asset index quintile largely confirm those of the overall analysis. Impact estimates tended to be slightly more positive and/or larger than for the overall sample. This is with the exception of the positive effect of T4 compared to T1, which appears not to have been present for the poorest 20%.

[11] This secular decline is likely attributable to a saturation in the population in response to intensive efforts for increased vaccination coverage in the last 15 years

Box 16: Proportion of recently pregnant women having received at least 2 doses of tetanus vaccine during pregnancy



All	-0.027
Poor	0.026

All	-0.068*	0.054	0.131**
Poor	-0.013	0.013	0.099

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	-0.025	-0.075	0.047	-0.085
Poor	0.040	0.029	0.054	-0.061

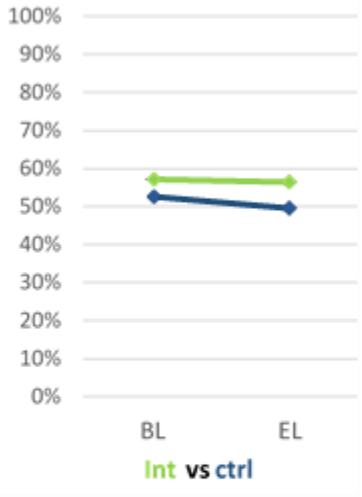
Indicator 17: Impact of PBF on the proportion of recently pregnant women having been offered HIV testing during pregnancy

Sample, indicator measurement and calculation. The sample was restricted to women who had ended a pregnancy within the 24 months prior to the interview date, irrespective of pregnancy outcome. The indicator was calculated as the proportion of the sample who reported to have been offered an HIV test in the course of their pregnancy.

Main results. Results pertaining to indicator 17 are displayed in **Box 17**. Overall, the proportion of women who had been offered HIV testing remained relatively stable at around 55%, with slightly higher rates in the intervention group. No impact of PBF compared to status quo (specification 1) was detected. Comparing the different intervention arms to status quo (specification 2), effect estimates for T1 and T3 are close to zero. The estimate for T2 is

Box 17: Proportion of recently pregnant women having been offered HIV testing during pregnancy

Specification 1 (quasi-experiment): Intervention (pooled) vs control



All	0.001
Poor	0.047

Specification 3 (experiment): T2, T3, and T4 vs T1

BL n=6,557
EL n=6,215

All	0.030	-0.012	0.007
Poor	0.085	0.041	-0.064

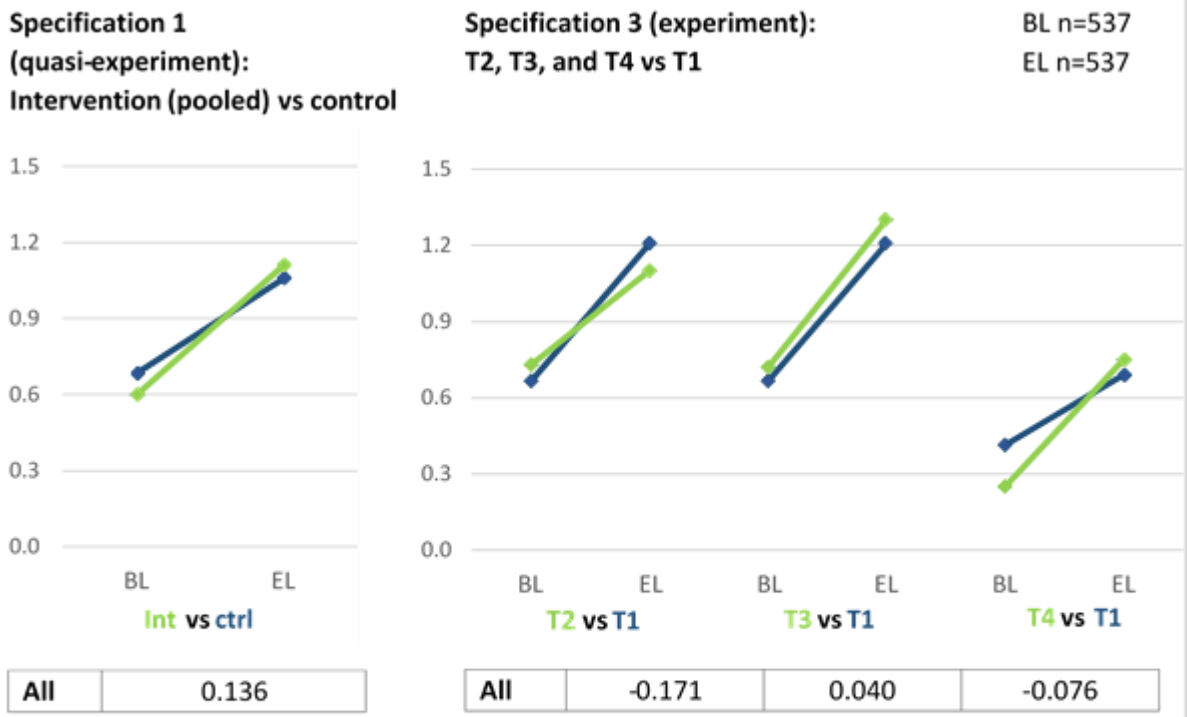
Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	0.025	0.073	0.033	-0.262**
Poor	0.061	0.163*	0.121	-0.255*

somewhat higher (+7.3 pp), but not statistically significant. For T4, results imply a strong negative effect compared to status quo, which however is again due to a strong decline in offer rates in Solenzo (-47 pp) and Nouna (-18 pp) overall as visible from the T4 vs T1 graph, calling into question the suitability of all control districts as a counterfactual in this particular case. In the experimental study component (specification 3), no additional benefit of the "adds-on" compared to standard PBF could be detected.

Stratified analysis on the poorest 20%. Results using the subsample in the lowest asset index quintile largely confirm those of the overall analysis. Effect estimates are positive with the exception of T4 and somewhat larger than in the overall analysis, particularly in T2 and T3 compared to status quo (+16% pp and +12% pp, respectively), the former even reaching statistical significance.

Box 18: Number of HIV-positive mothers who have completed prophylactic ARV treatment (SNIS)



All	0.136

All	-0.171	0.040	-0.076

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	0.172	-0.005	0.205	0.124

Indicator 18: Impact of PBF on the number of HIV-positive mothers who have completed prophylactic ARV treatment (SNIS)

Indicator measurement and calculation. This indicator was based on data from the routine health information system, which registers on a monthly basis the number of women who have completed ARV for PMTCT in each health facility. As described in 2.4, the indicator reflects the average number of women with complete prophylactic ARV treatment over a six-month interval at baseline and endline per health facility. Accordingly, effect estimates are to be interpreted as absolute change in case numbers. As SNIS data are not stratified by patient socio-economic status, no stratified analysis was possible for this indicator.

Results. Results pertaining to indicator 18 are displayed in **Box 18**. Overall, the average number of complete prophylactic ARV treatments remained very low between baseline and endline, increasing from 0.6 to 1.1 over a six-month period, with little variation between the intervention and control groups. None of the DID analyses indicate any impact of PBF on the number of prophylactic treatments completed.

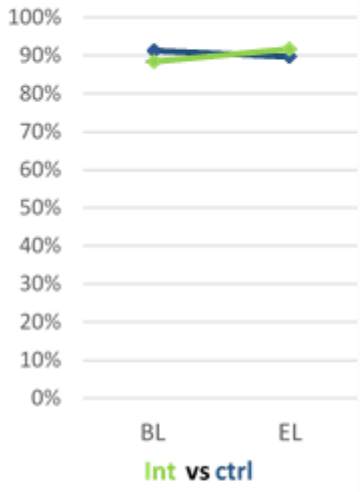
Indicator 19: Impact of PBF on the proportion of recently pregnant women who have delivered in a formal health facility

Sample, indicator measurement and calculation. The sample was restricted to women who had ended a pregnancy within the 24 months prior to the survey and whose pregnancy outcome had been a live birth or still birth. Women whose pregnancy had ended in an abortion or a miscarriage were not included in the computation of this indicator. The indicator was calculated as the proportion of the sample having delivered in a formal health facility.

Main results. Results pertaining to indicator 19 are displayed in **Box 19**. Overall, the proportion of women having delivered in a formal health facility was high at around 90%. The comparison of PBF to status quo indicates a positive intervention effect of +4.4 pp overall

Box 19: Proportion of recently pregnant women who have delivered in a formal health facility

Specification 1 (quasi-experiment): Intervention (pooled) vs control



All	0.044*
Poor	0.022

Specification 3 (experiment): T2, T3, and T4 vs T1

BL n=6,511
EL n=6,220

All	-0.033	-0.038*	0.008
Poor	-0.117**	-0.118**	-0.014

Specification 2 (quasi-experiment): Intervention vs control

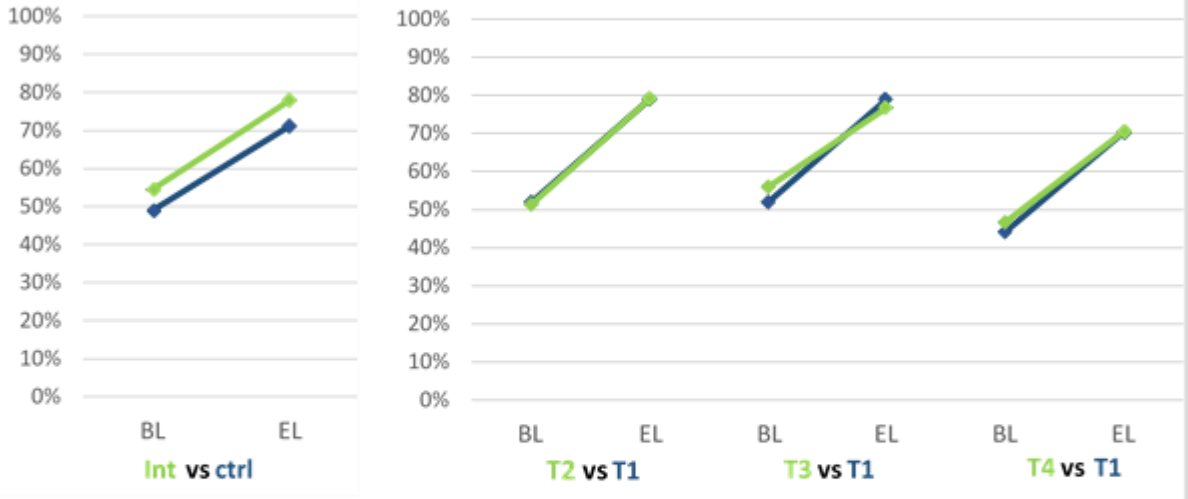
	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	0.052**	0.047*	0.042**	0.011
Poor	0.057	0.007	0.007	-0.051

Box 20: Proportion of recently pregnant women with at least one PNC visit within 6 weeks after delivery

Specification 1 (quasi-experiment): Intervention (pooled) vs control

Specification 3 (experiment): T2, T3, and T4 vs T1

BL n=6,488
EL n=6,194



All	0.030
Poor	-0.006

All	0.012	-0.067*	-0.016
Poor	-0.041	-0.103	-0.241**

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	0.024	0.083	0.001	0.019
Poor	0.018	0.012	-0.050	-0.043

(specification 1) as well as in intervention arms T2, T3, and T4 (+5.2 pp, +4.7 pp, and 4.2 pp, respectively). Results from the experimental study component (specification 3) confirm a slightly more positive change in T1 compared to T2 and T3. For T4, neither an effect compared to controls nor to T1 could be detected.

Stratified analysis on the poorest 20%. Among the poorest 20%, effect estimates for the comparison of PBF to controls are positive except for T4, but do not reach statistical significance. As in the overall sample but to a larger extent, results show a negative incremental effect of T2 and T3 above T1 (-11.7 pp and -11.8 pp, respectively).

Indicator 20: Impact of PBF on the proportion of recently pregnant women with at least one PNC visit within 6 weeks after delivery

Sample, indicator measurement and calculation. The sample was restricted to women who had ended a pregnancy within the 24 months prior to the survey and whose pregnancy outcome had been a live birth or still birth, regardless of place of delivery. The indicator was calculated as the proportion of the sample who had had at least one postnatal consultation visit in a formal health facility within six weeks after delivery.

Main results. Results pertaining to indicator 20 are displayed in **Box 20**. The proportion of women with at least one PNC visit increased from an average of 53% at baseline to 76% at endline, with slightly lower proportion but similar increase in the control group. No impact of PBF compared to status quo could be detected (specifications 1 and 2). The experimental study component (specification 3) shows no difference between T1 and T2 and T4. In T3, in contrast, the incline in PNC utilization was somewhat lower (-7 pp).

Stratified analysis on the poorest 20%. No impact of PBF compared to status quo was apparent for the poorest 20% either. Unlike in the overall sample, there was a negative incremental effect of T4 compared to the basic T1 of -24 pp in the subsample of the poorest.

Indicator 21: Impact of PBF on the proportion of recently pregnant women with at least three PNC visits within 6 weeks after delivery

Sample, indicator measurement and calculation. Sample and indicator measurement were identical to indicator 20. The indicator was calculated as the proportion of the sample who had had at least three postnatal consultation visits in a formal health facility within six weeks after delivery.

Main results. Results pertaining to indicator 21 are displayed in **Box 21**. The proportion of women with three PNC visits in the first six weeks after delivery increased from 5% to 14% between baseline and endline in the intervention group, and from 7% to 11% in the control group. The DID analysis (specifications 1 and 2) confirm a positive intervention effect compared to status quo of +6.6 pp, particularly in intervention arms T1 and T2 (+7.2 and +8.0 pp, respectively), and somewhat less and not statistically significant in T3 and T4. There is no evidence of an added benefit of the “adds-on” compared to the standard T1 in the experimental study component (specification 3).

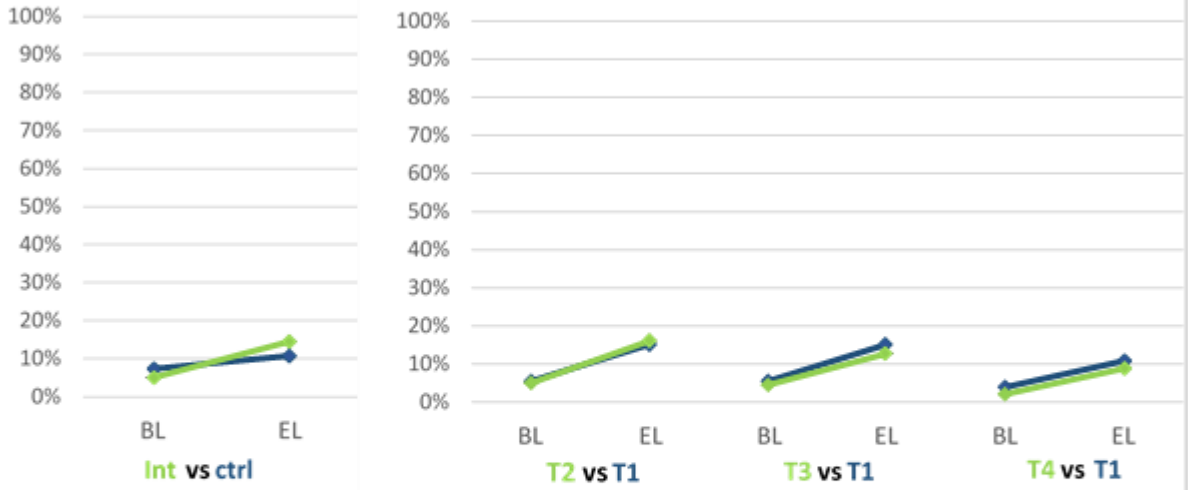
Stratified analysis on the poorest 20%. Unlike in the overall sample, no impact of PBF compared to status quo could be detected for the poorest 20%. Estimates are positive, but of small magnitude with the exception of T3 (+5.9 pp, not significant). No added benefit of T2, T3, and T4 beyond T1 is apparent.

Box 21: Proportion of recently pregnant women with at least three PNC visits within 6 weeks after delivery

Specification 1 (quasi-experiment): Intervention (pooled) vs control

Specification 3 (experiment): T2, T3, and T4 vs T1

BL n=6,488
EL n=6,194



All	0.066*
Poor	0.041

All	0.018	-0.010	-0.022
Poor	0.052	0.076	-0.069

Specification 2 (quasi-experiment): Intervention vs control

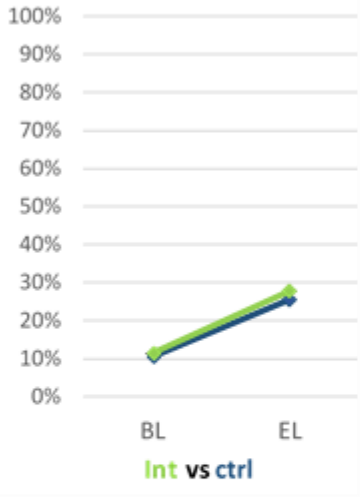
	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	0.072**	0.080*	0.051	0.040
Poor	0.041	0.037	0.059	0.019

Indicator 22: Impact pf PBF on the proportion of non-pregnant women aged 15-49 who use modern family planning methods

Sample, indicator measurement and calculation. For this indicator, the sample was restricted to all women who had responded to the women's module of the household survey and were not pregnant at the time of the survey, irrespective of whether they had ended a pregnancy in the 24 months prior to the interview date. The indicator was calculated as the proportion of the sample who indicated using any modern methods of family planning (including sterilization, IUDs, hormonal methods (injections, implants, pill), condoms, condoms, diaphragms, or mousse/gel).

Box 22: Proportion of non-pregnant women aged 15-49 who use modern family planning methods

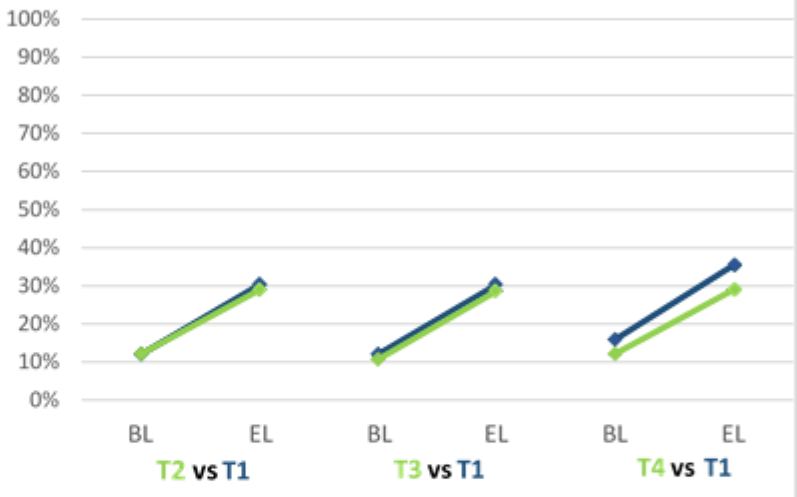
Specification 1 (quasi-experiment): Intervention (pooled) vs control



All	0.019
Poor	0.076*

Specification 3 (experiment): T2, T3, and T4 vs T1

BL n=8,643
EL n=8,942



All	-0.021	-0.011	-0.031
Poor	-0.022	-0.127**	0.023

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	0.010	0.026	0.035	0.012
Poor	0.071	0.113**	0.010	0.142***

Main results. Results pertaining to indicator 22 are displayed in **Box 22**. Utilization rates of modern family planning methods increased from 11% at baseline to 27% at endline, with almost no difference in the intervention and control groups. No impact of PBF compared to status quo is apparent, neither overall, nor for the different intervention arms (specifications 1 and 2). There was no evidence of an added benefit of the "adds-on" compared to the standard T1 (specification 3).

Stratified analysis on the poorest 20%. Unlike what observed for the entire sample, a positive intervention impact on the use of modern family planning could be detected among the poorest 20% overall (+7.6 pp) as well as in intervention arms T1, T2, and T4, but not in intervention arm T3.

Summary: Impact of PBF on the utilization of reproductive health care services

Table 11 and **Table 12** summarize the impact estimates for the nine indicators pertaining to the utilization of reproductive health care services for the full sample as well as the subsample of the poorest 20%, respectively. Positive and statistically significant impact estimates are marked in green, negative and significant impact estimates in red. Cells not marked in color contain estimates that did not reach statistical significance.

Table 11: Summary of results pertaining to the impact of PBF on the utilization of reproductive health care services (full sample)

	Quasi-experiment					Experiment		
	PBF vs control	T1 vs control	T2 vs control	T3 vs control	T4 vs control	T2 vs T1	T3 vs T1	T4 vs T1
14: Four ANC visits	-0.004	-0.011	0.038	0.046	-0.132***	-0.026	-0.018	-0.031
15: ANC in 1st trimester	0.057	0.050	0.069	0.097	0.003	-0.036	-0.007	-0.064
16: Tetanus vaccination in preg.	-0.027	-0.025	-0.075	0.047	-0.085	-0.068*	0.054	0.131**
17: HIV test offer in pregnancy	0.001	0.025	0.073	0.033	-0.262**	0.030	-0.012	0.007
18: PMTCT (SNIS)	0.136	0.172	-0.005	0.205	0.124	-0.171	0.040	-0.076
19: Facility-based delivery	0.044*	0.052**	0.047*	0.042**	0.011	-0.033	-0.038*	0.008
20: One PNC visit	0.030	0.024	0.083	0.001	0.019	0.012	-0.067*	-0.016
21: Three PNC visits	0.066*	0.072**	0.080*	0.051	0.040	0.018	-0.010	-0.022
22: Modern family planning	0.019	0.010	0.026	0.035	0.012	-0.021	-0.011	-0.031

Note: * p < 0.1, ** p < 0.05, *** p < 0.01. All effect estimates pertain to absolute (as opposed to relative) change. Effect estimates for all indicators but 18 can be converted to percentages and reflect percentage point changes. Indicator 18 is based on SNIS data, so estimates correspond to absolute change in the average half-yearly number of patients per facility attributable to the intervention.

Table 12: Summary of results pertaining to the impact of PBF on the utilization of reproductive health care services (poorest 20%)

	Quasi-experiment					Experiment		
	PBF vs control	T1 vs control	T2 vs control	T3 vs control	T4 vs control	T2 vs T1	T3 vs T1	T4 vs T1
14: Four ANC visits	0.037	0.031	0.111	0.125**	-0.162**	-0.038	-0.026	0.020
15: ANC in 1st trimester	0.031	0.017	0.023	0.057	0.048	-0.071	-0.037	0.073
16: Tetanus vaccination in preg.	0.026	0.040	0.029	0.054	-0.061	-0.013	0.013	0.099
17: HIV test offer in pregnancy	0.047	0.061	0.163*	0.121	-0.255*	0.085	0.041	-0.064
18: PMTCT (SNIS)								
19: Facility-based delivery	0.022	0.057	0.007	0.007	-0.051	-0.117**	-0.118**	-0.014
20: One PNC visit	-0.006	0.018	0.012	-0.050	-0.043	-0.041	-0.103	-0.241**
21: Three PNC visits	0.041	0.041	0.037	0.059	0.019	0.052	0.076	-0.069
22: Modern family planning	0.076*	0.071	0.113**	0.010	0.142***	-0.022	-0.127**	0.023

Note: * p < 0.1, ** p < 0.05, *** p < 0.01. All effect estimates pertain to absolute (as opposed to relative) change. Effect estimates for all indicators but 18 can be converted to percentages and reflect percentage point changes. Indicator 18 is based on SNIS data, so estimates correspond to absolute change in the average half-yearly number of patients per facility attributable to the intervention.

3.6. Impact of PBF on the utilization of preventive child health services

In this section, results pertaining to the impact of PBF on the utilization of preventive child health care services are presented. Specific indicators include:

23. Proportion of children aged 12-23 months who are fully immunized (primary data); number of children aged 0-11 months fully immunized (SNIS)
24. Proportion of children aged 0-11 months who have participated in growth monitoring in last 6 months (primary data); number of new growth monitoring visits of children aged 0-11 months (SNIS)
25. Proportion of children aged 12-23 months who have participated in growth monitoring in last 6 months

Primary data for the three indicators were extracted from the household survey module for children under 5. We further used routine data to check for robustness of the results in light of the potential sample bias discussed in 2.5. For indicator 25, the equivalent routine data were unfortunately too incomplete to use.

Indicator 23: Impact of PBF on the proportion of children aged 12-23 months who are fully immunized / number of children aged 0-11 months fully immunized (SNIS)

Sample, indicator measurement and calculation. For this indicator, the sample was restricted to all children aged 12-23 months. The indicator was calculated as the proportion of the sample who had received all nine basic vaccinations to be given during the first year of life according to the national vaccination calendar[12], including one dose of BCG, three doses of OPV, three doses of pentavalent, one dose of measles, and one dose of yellow fever vaccine. Note that we did not consider the timing of the respective vaccinations, but rather used vaccination rates among children aged 12-23 months as a proxy for adherence to the vaccination guidelines, which however pertain to children under the age of one.

The corresponding SNIS indicator is a count of the number of children with timely completion of the basic vaccination cycle per facility in a six-month interval. Effect estimates are to be interpreted as absolute change in case numbers, accordingly. Note that unlike for all other SNIS indicators, we used data from October 2015 to March 2016 (rather than Oct 16 - Mar 17) as the endline period because of a nation-wide lack of data on this indicator in the second half of 2016. Hence, the endline period for the primary and routine data are not aligned, accordingly.

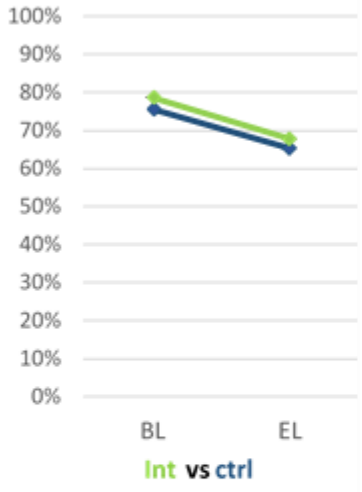
Main results. Results pertaining to indicator 23a (primary data) are displayed in **Box 23a**. Overall, the proportion of fully immunized children decreased from overall 78% to 67% between baseline and endline, with only very small differences between the intervention and control group[13]. No impact of PBF compared to status quo could be detected, neither overall

[12]http://www.nationalplanningcycles.org/sites/default/files/country_docs/Burkina%20Faso/ppac_2011_2015_dpv_revise_30_aout_2012_1.pdf

[13] Note that there appears to have been a nationwide shortage of vaccines in the latter half of 2016, which likely contributed to this secular decline

Box 23a: Proportion of children aged 12-23 months who are fully immunized

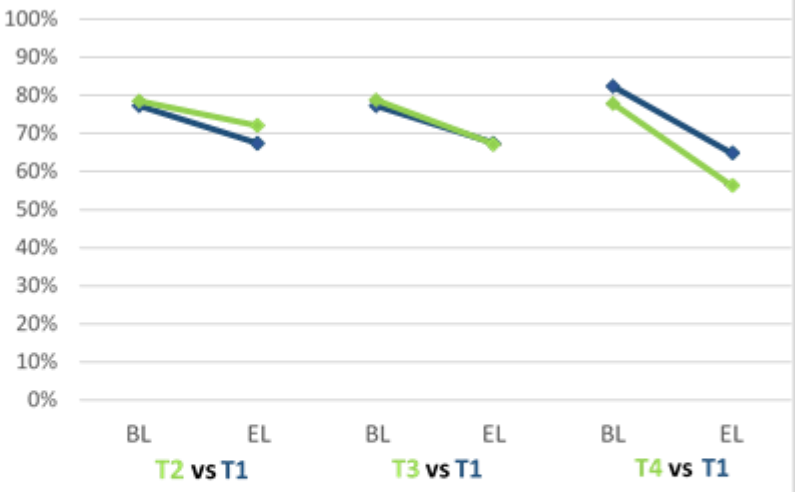
Specification 1 (quasi-experiment): Intervention (pooled) vs control



All	-0.001
Poor	-0.022

Specification 3 (experiment): T2, T3, and T4 vs T1

BL n=3,262
EL n=2,988



All	0.018	-0.026	0.024
Poor	-0.077	-0.122	-0.256

Specification 2 (quasi-experiment): Intervention vs control

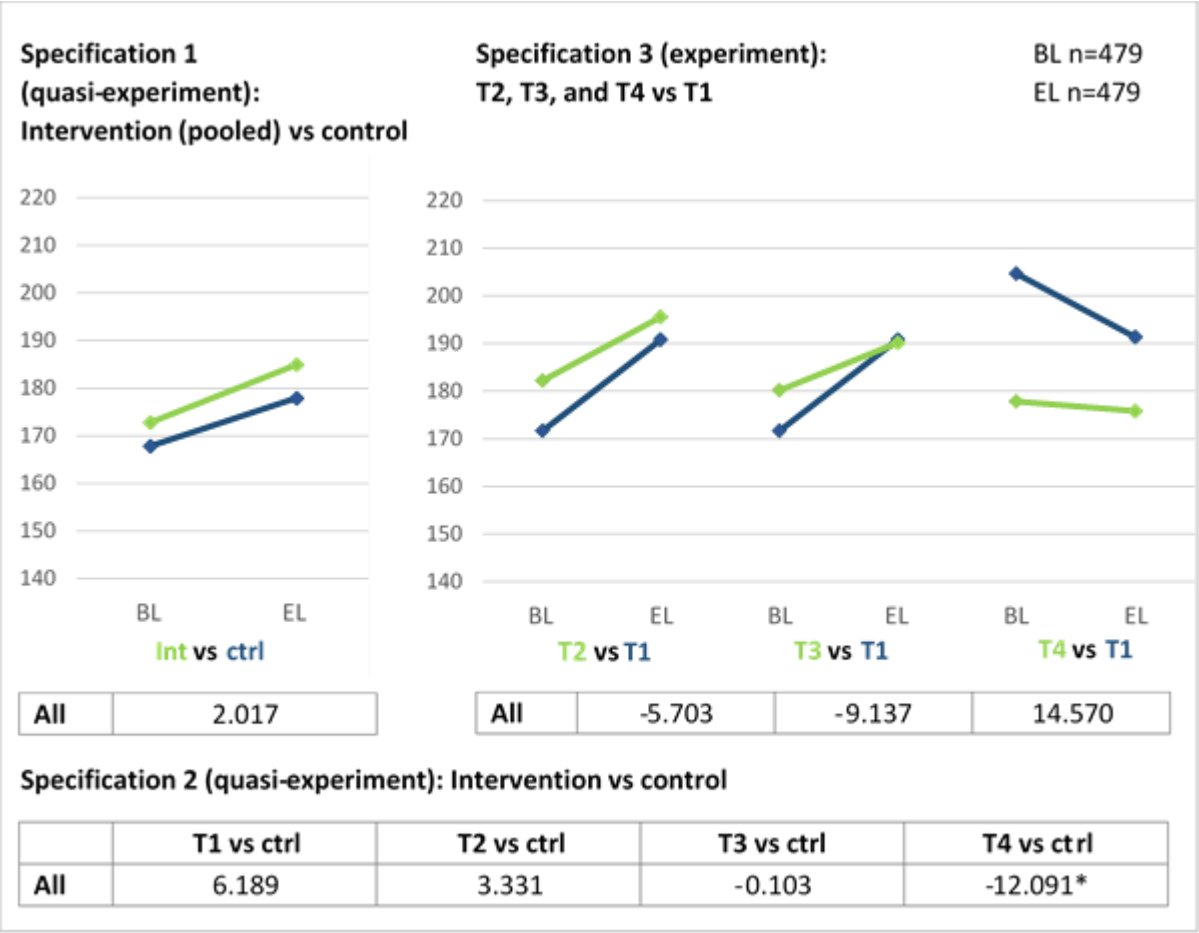
	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	0.013	0.040	-0.007	-0.129
Poor	0.061	-0.014	-0.055	-0.350**

(specification 1), nor for intervention arms T1, T2, and T3 (specification 2). The comparatively large though statistically not significant negative effect for T4 compared to controls is at least in part attributable to a particularly strong decline in vaccination rates in the Nouna district (-40 pp). We found no evidence of an added benefit of the "add-on" compared to the standard T1 (specification 3).

Results based on routine data (indicator 23b) can be found in **Box 23b**. At facility level, there was a very slight increase in case numbers between baseline and endline of on average about 11 patients per facility in a six-month interval. Results from the DID based on secondary data confirm the results obtained using primary data, showing no impact of PBF compared to status quo.

Stratified analysis on the poorest 20%. The results of the analysis on the poorest 20% of the sample largely mirror those of the overall analysis, with the exception of a particularly strong negative effect of T4 both compared to controls (only partly attributable to the trend in Nouna) and to T1 in the experimental study component.

Box 23b: Number of children aged 0-11 months fully immunized (SNIS)



All	2.017

All	-5.703	-9.137	14.570

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	6.189	3.331	-0.103	-12.091*

Indicator 24: Impact of PBF on the proportion of children aged 0-11 months who have participated in growth monitoring in last 6 months / number of new growth monitoring visits of children aged 0-11 months (SNIS)

Sample, indicator measurement and calculation. For this indicator, the sample was restricted to all children aged 0-11 months. The indicator was calculated as the proportion of the sample having been measured to determine their nutritional status in the 6 months prior to the survey, either in a health facility or as part of an outreach or community health worker activity or campaign.

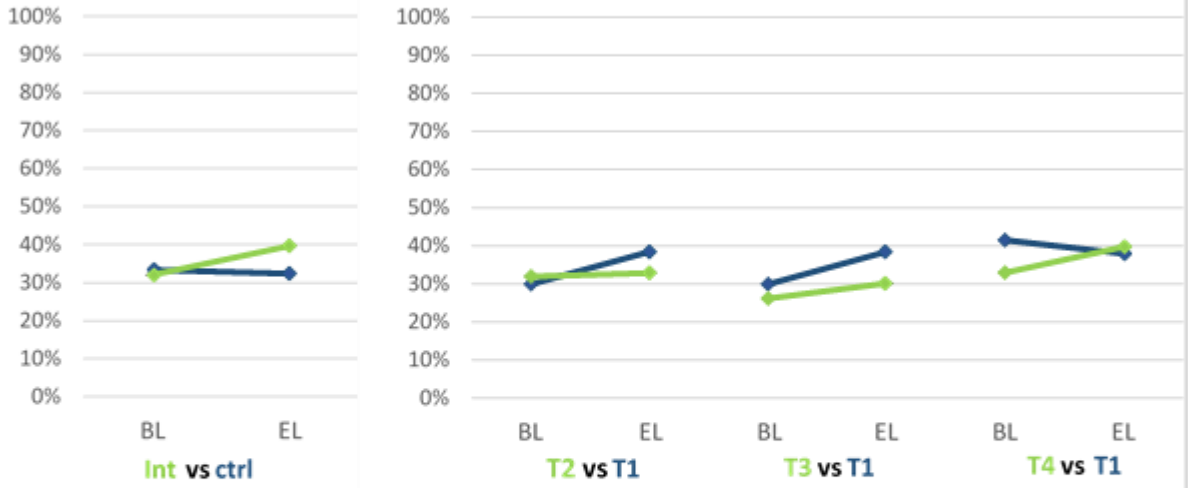
The corresponding SNIS indicator is a count of the number of children aged 0-11 newly registered for growth monitoring ("consultation du nourrisson sain") in each health facility in a six-month interval. Accordingly, effect estimates are to be interpreted as absolute change in case numbers.

Box 24a: Proportion of children aged 0-11 months who have participated in growth monitoring in last 6 months

Specification 1 (quasi-experiment): Intervention (pooled) vs control

Specification 3 (experiment): T2, T3, and T4 vs T1

BL n=3,414
EL n=3,737



All	0.079
Poor	0.018

All	-0.070	-0.051	0.069
Poor	-0.197*	-0.192*	0.158

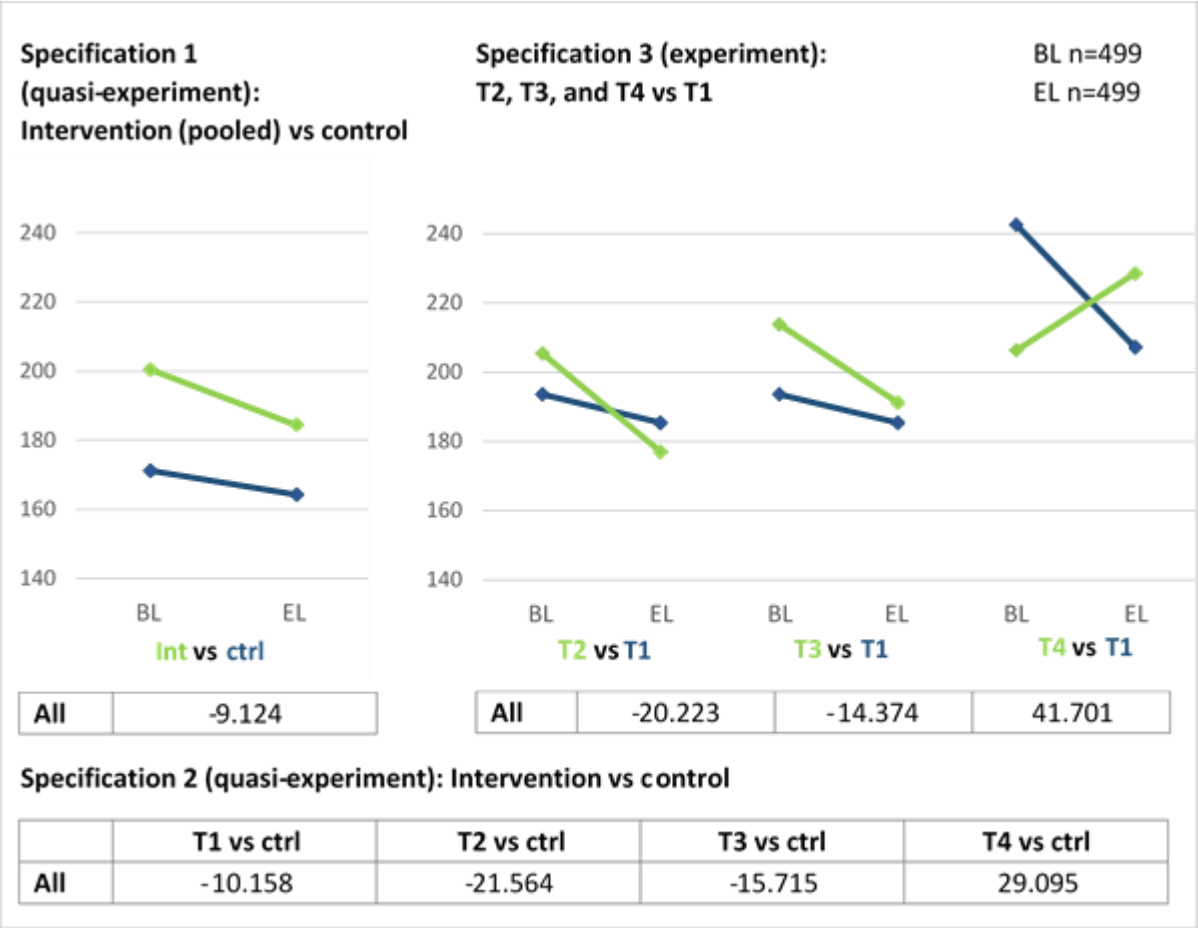
Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	0.128	0.010	0.027	0.082
Poor	0.123	-0.110	-0.120	-0.007

Main results. Results pertaining to indicator 24a (primary data) are displayed in **Box 24a**. Overall, growth monitoring utilization rates increased from 32% at baseline in both study groups to 40% in the intervention group, while remaining stable in the control group. The corresponding impact estimate (+7.9 pp; specification 1) is not statistically significant, however, largely driven by intervention arms T1 and T4 (specification 2). This is also reflected in the results of the experimental component (specification 3), with T2 and T3 performing slightly worse than T1, and T4 slightly better. None of the coefficients are statistically significantly different from zero, however, so that there is no evidence of an added benefit of the "adds-on" beyond the standard T1 in regards to growth monitoring among children under the age of one.

Results based on routine data (indicator 24b) can be found in **Box 24b**. Routine data showed a very slight decline in case numbers at facility level (-7% relative to baseline), unlike what the primary data suggests. Results of the DID analyses, however, confirm the results obtained using primary data, showing no impact of PBF compared to status quo and indicating similar

Box 24b: Number of new growth monitoring visits of children aged 0-11 months (SNIS)



All	-9.124

All	-20.223	-14.374	41.701

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	-10.158	-21.564	-15.715	29.095

patterns in regards to the performance of the different intervention arms. Restricting the observation period to before the on-set of the gratuité policy leads to similar results.

Stratified analysis on the poorest 20%. Unlike for the overall sample, there is no indication of a positive impact of PBF in the subsample of the poorest 20%. However, the overall null effect masks variation between the different intervention arms, with T1 appearing to have produced positive impact (not significant) and the opposite being observed in T2 and T3.

Indicator 25: Impact of PBF on the proportion of children aged 12-23 months who have participated in growth monitoring in last 6 months

Sample, indicator measurement and calculation. For this indicator, the sample was restricted to all children aged 12-23 months. The indicator was measured and calculated analogous to indicator 24. No SNIS data corresponding to the primary data were available for this indicator.

Box 25: Proportion of children aged 12-23 months who have participated in growth monitoring in last 6 months

Specification 1 (quasi-experiment): Intervention (pooled) vs control

Specification 3 (experiment): T2, T3, and T4 vs T1

BL n=3,232
EL n=2,945

All	-0.110
Poor	-0.112

All	-0.036	-0.039	-0.033
Poor	-0.090	-0.119	-0.077

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	-0.060	-0.183**	-0.190*	-0.050
Poor	-0.026	-0.213	-0.238	-0.009

Main results. Results pertaining to indicator 25 are displayed in **Box 25**. For children aged 12-23 months, growth monitoring rates stayed stable at 31% in the intervention group, while they increased from 20% to 31% in the control group between baseline and endline. The corresponding DID analyses (specification 1 and 2) indicates a negative intervention effect compared to status quo of -11 pp overall (not significant), particularly driven by intervention arms T2 (-18.3 pp) and T3 (-19.0 pp). Interestingly, this is only somewhat reflected in the experimental component of the study (specification 3), reflecting variation between districts.

Stratified analysis on the poorest 20%. Results from the subsample of the poorest 20% mirror those of the analysis on the entire sample.

Summary: Impact of PBF on the utilization of preventive child health services

Table 13 and **Table 14** summarize impact estimates for the indicators pertaining to the utilization of preventive and routine monitoring child health services for the full sample as well as the subsample of the poorest 20%, respectively. Positive and statistically significant impact estimates are marked in green, negative and significant impact estimates in red. Cells not marked in color contain estimates that did not reach statistical significance.

Table 13: Summary of results pertaining to the impact of PBF on the utilization of preventive child health services (full sample)

	Quasi-experiment					Experiment		
	PBF vs control	T1 vs control	T2 vs control	T3 vs control	T4 vs control	T2 vs T1	T3 vs T1	T4 vs T1
23a: Children fully immunized (primary data)	-0.001	0.013	0.040	-0.007	-0.129	0.018	-0.026	0.024
23b: Children fully immunized (SNIS)	2.017	6.189	3.331	-0.103	-12.091*	-5.703	-9.137	14.570
24a: Growth monitoring children 0-11 months (primary data)	0.079	0.128	0.010	0.027	0.082	-0.070	-0.051	0.069
24b: Growth monitoring children 0-11 months (SNIS)	-9.124	-10.158	-21.564	-15.715	29.095	-20.223	-14.374	41.701
25: Growth monitoring children 12-23 months	-0.110	-0.060	-0.183**	-0.190*	-0.050	-0.036	-0.039	-0.033

Note: * p < 0.1, ** p < 0.05, *** p < 0.01. All effect estimates pertain to absolute (as opposed to relative) change. Effect estimates for primary data indicators can be converted to percentages and reflect percentage point changes. Effect estimates for SNIS indicators correspond to absolute change in the average half-yearly number of patients per facility attributable to the intervention.

Table 14: Summary of results pertaining to the impact of PBF on the utilization of preventive child health services (poorest 20%)

	Quasi-experiment					Experiment		
	PBF vs control	T1 vs control	T2 vs control	T3 vs control	T4 vs control	T2 vs T1	T3 vs T1	T4 vs T1
23a: Children fully immunized (primary data)	-0.022	0.061	-0.014	-0.055	-0.350**	-0.077	-0.122	-0.256
23b: Children fully immunized (SNIS)								
24a: Growth monitoring children 0-11 months (primary data)	0.018	0.123	-0.110	-0.120	-0.007	-0.197*	-0.192*	0.158
24b: Growth monitoring children 0-11 months (SNIS)								
25: Growth monitoring children 12-23 months	-0.112	-0.026	-0.213	-0.238	-0.009	-0.090	-0.119	-0.077

Note: * p < 0.1, ** p < 0.05, *** p < 0.01. All effect estimates pertain to absolute (as opposed to relative) change. Effect estimates for primary data indicators can be converted to percentages and reflect percentage point changes. Effect estimates for SNIS indicators correspond to absolute change in the average half-yearly number of patients per facility attributable to the intervention.

3.7. Impact of PBF on the utilization of curative health care services

In this section, result pertaining to the impact of PBF on the utilization of curative health care services are presented. Specific indicators include:

26. Number of patients under age 5 having sought curative services (SNIS)
27. Number of patients age 5 or older having sought curative services (SNIS)

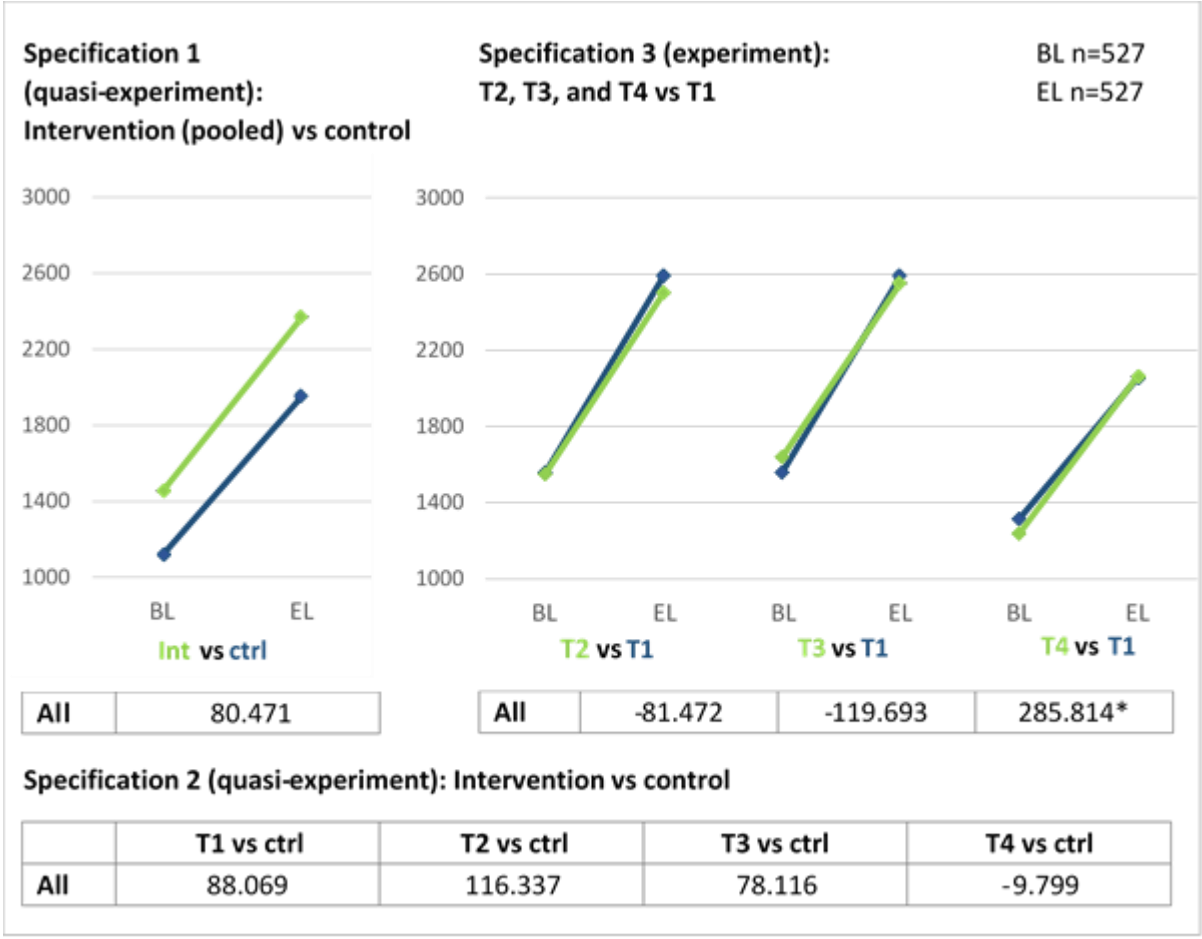
Data for the calculation of the two indicators were extracted from the routine health information system (SNIS), specifically data for October 2013 to March 2014 (baseline) and for October 2016 to March 2017 (endline). We did not work with primary data for these two indicators given that the number of respondents reporting an acute illness episode in the household survey was simply too low to allow for meaningful analysis across intervention arms.

Indicator 26: Impact of PBF on the number of patients under age 5 having sought curative services (SNIS)

Indicator measurement and calculation. The indicator is a count of the number of children having sought care for curative services per facility in a six-month interval. Accordingly, effect estimates are to be interpreted as absolute change in case numbers, accordingly. As SNIS data are not stratified by patient socio-economic status, no stratified analysis was possible for this indicator.

Results. Results pertaining to indicator 26 are displayed in **Box 26**. Overall, numbers of patients under 5 increased by around 65% between baseline and endline, from an average of around 1380 per facility between October ’13 to March ’14 to an average of around 2280 between October ’16 to March ’17. Average patient numbers were somewhat higher in intervention facilities compared to control facilities at both baseline and endline. DID estimates for the impact of PBF compared to status quo (specifications 1 and 2) are all positive but not statistically significant, with the exception of T4, which is negative but relatively close to zero, considering the scale of the indicator. Compared to the basic T1, change appeared to have been less positive in T2 and T3, although impact estimates are far from statistical significance (specification 3), indicating no additional benefit of targeting for utilization of curative consultation among children under 5. In T4 facilities, in contrast, results indicate a stronger increase in patient numbers than in T1 facilities. Additional analysis using October 2015 to March 2016 as the endline interval (i.e. before the introduction of the gratuité policy) resulted in negative but statistically insignificant impact estimates for the comparisons of PBF to status quo (specifications 1 and 2), and similar results regarding the differential performance of T2, T3, and T4 over and above the standard T1.

Box 26: Number of patients under age 5 having sought curative services (SNIS)



All	80.471			
All	-81.472	-119.693	285.814*	

Specification 2 (quasi-experiment): Intervention vs control

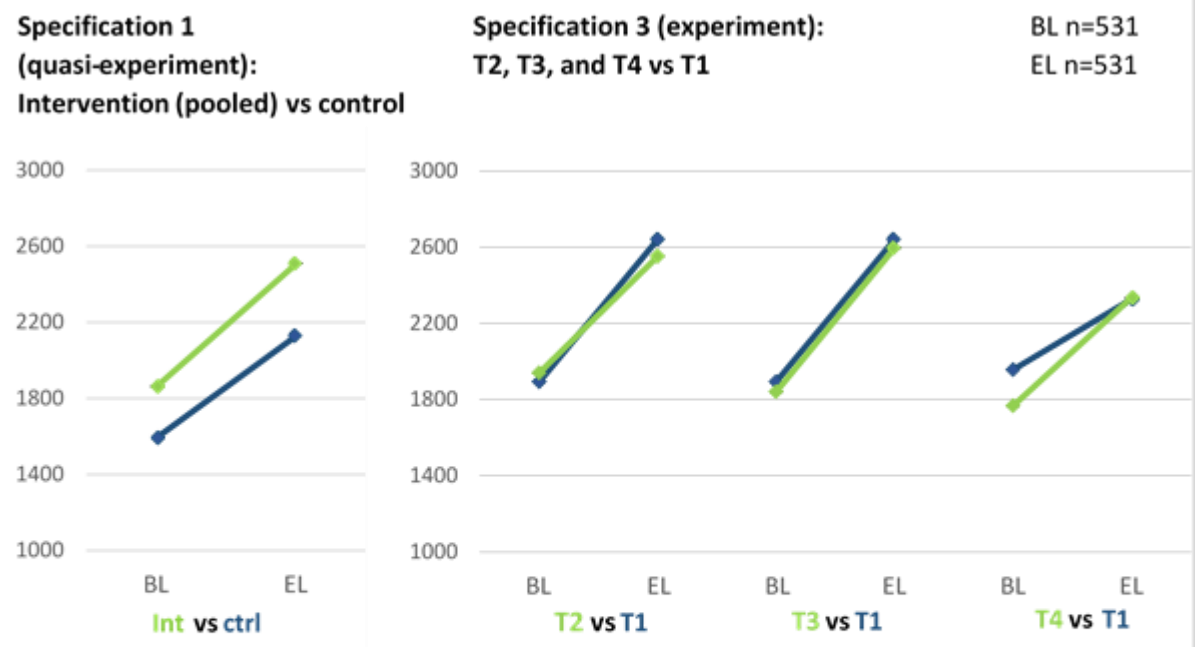
	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	88.069	116.337	78.116	-9.799

Indicator 27: Impact of PBF on the number of patients age 5 or older having sought curative services (SNIS)

Indicator measurement and calculation. The indicator is a count of the number of children aged 5 or older as well as adults having sought care for curative services per facility in a six-month interval. Children over the age of 5 were pooled with adults in alignment with the PBF indicator 1 (**Table 1**), which also pools all age groups except for young children under 5. Effect estimates are to be interpreted as absolute change in case numbers.

Results. Results pertaining to indicator 27 are displayed in **Box 27**. Overall, numbers of patients aged 5 or older increased by around 35% between baseline and endline, from an average of around 1800 per facility between October ’13 to March ’14 to an average of around 2425 between October ’16 to March ’17. Average patient numbers were somewhat higher in intervention facilities compared to control facilities at both baseline and endline, and increased somewhat more in intervention facilities. DID estimates for the impact of PBF compared to status quo (specifications 1 and 2) are all positive accordingly, but not statistically significant with the exception of T3. Within the experimental study areas (specification 3), there is no evidence of differential changes in T2 and T3 compared to T1.

Box 27: Number of patients age 5 or older having sought curative services (SNIS)



All	109.608

All	-134.890	8.296	396.714*

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	94.186	77.098	220.284*	32.573

In T4 facilities, however, the positive change is more pronounced than in the corresponding T1 facilities. Restricting the observation period to before the on-set of the gratuité policy leads to similar but more pronounced results. The overall impact estimate for the comparison of PBF to status quo as well as those for the different intervention arms are statistically significant, showing positive impact of PBF in the pre-gratuité period.

Summary: Impact of PBF on the utilization of curative health care services

Table 15 summarizes impact estimates for the two indicators pertaining to the utilization of curative health care services. Positive and statistically significant impact estimates are marked in green, negative and significant impact estimates in red. Cells not marked in color contain estimates that did not reach statistical significance.

Table 15: Summary of results pertaining to the impact of PBF on the utilization of curative health care services

	Quasi-experiment					Experiment		
	PBF vs control	T1 vs control	T2 vs control	T3 vs control	T4 vs control	T2 vs T1	T3 vs T1	T4 vs T1
26: Curative consultation U5	81	88	116	78	-10	-82	-120	286*
27: Curative consultations 5+	110	94	77	220*	33	-135	8	397*

Note: * p < 0.1, ** p < 0.05, *** p < 0.01. All effect estimates pertain to absolute (as opposed to relative) change. As the indicators are based on SNIS data, estimates correspond to absolute change in the average half-yearly number of patients per facility attributable to the intervention.

3.8. Impact of PBF on population health indicators

In this section, results pertaining to the impact of PBF on the following selected population health indicators as presented:

28. Proportion of children aged 0-59 months who are severely stunted
29. Proportion of children aged 0-59 months with severe acute malnutrition
30. Proportion of children aged 6-59 months with anemia
31. Proportion of women aged 15-49 years with anemia

Data for the calculation of the four population health indicators were extracted from the anthropometry and biomarker module of the household survey administered to children under 5 and women of reproductive age. Unlike for all previous indicators, for indicators in this section, negative impact estimates indicate positive change, as they imply a reduction in illness burden.

It is important to note that the data collection periods were not fully aligned, with baseline data having been collected from October to March, and endline data from April to June. Although we did everything possible to avoid collecting data during the midst of the rainy season, we cannot exclude that some seasonality effects impacted the secular trends visible in the data to some extent, accordingly. As the differences in data collection periods were consistent across intervention and control districts, however, this does not affect the interpretability of the impact estimates presented below.

Indicator 28: Impact of PBF on the proportion of children aged 0-59 months who are severely stunted

Sample, indicator measurement and calculation. For this indicator, the sample was restricted to all children aged 0-59 months for whom height was measured. Children were considered

as severely stunted if their height-for-age z-scores were below three standard deviations from the reference median based on the 2006 WHO child growth standards[14].

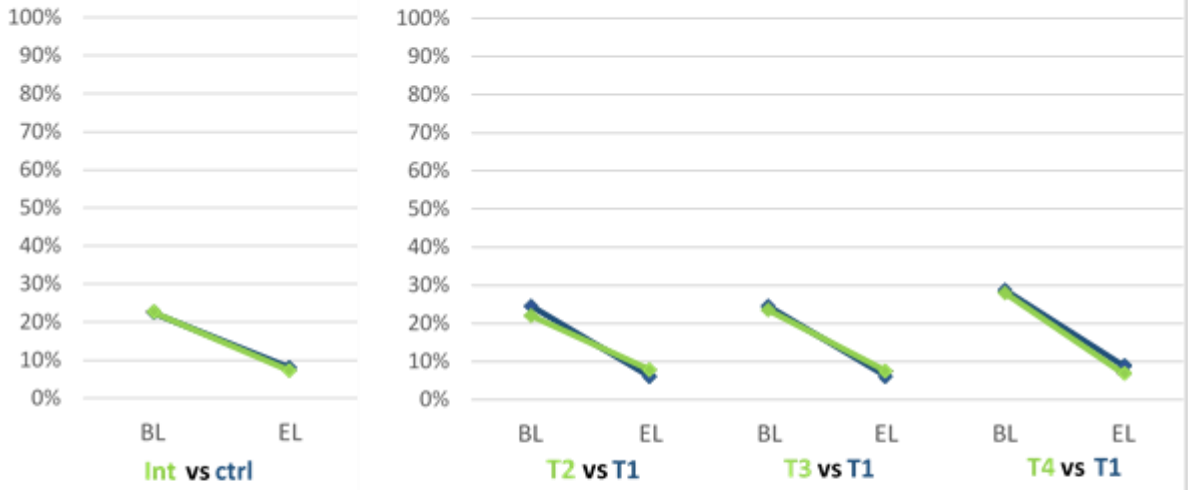
Main results. Results pertaining to indicator 28 are displayed in **Box 28**. The proportion of severely stunted children reduced from an 23% to 7% between baseline and endline in both the intervention and control groups. The corresponding impact estimates (specification 1 and 2) indicate no impact of PBF compared to status quo. This is with the exception of a significant effect estimate for T4, which is largely attributable to an above-average decline in stunting rates in the Nouna and Solenzo districts The experimental study component (specification 3) indicates a small incremental effect of T2 over and above T1, in that stunting rates did not decrease quite as much in T2 as in T1.

Box 28: Proportion of children aged 0-59 months who are severely stunted

Specification 1 (quasi-experiment): Intervention (pooled) vs control

Specification 3 (experiment): T2, T3, and T4 vs T1

BL n=9,673
EL n=12,604



All	-0.012
Poor	0.030

All	0.041**	0.022	-0.018
Poor	0.023	-0.058	0.084

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	-0.001	-0.002	-0.021	-0.061***
Poor	0.052	0.061	-0.026	-0.006

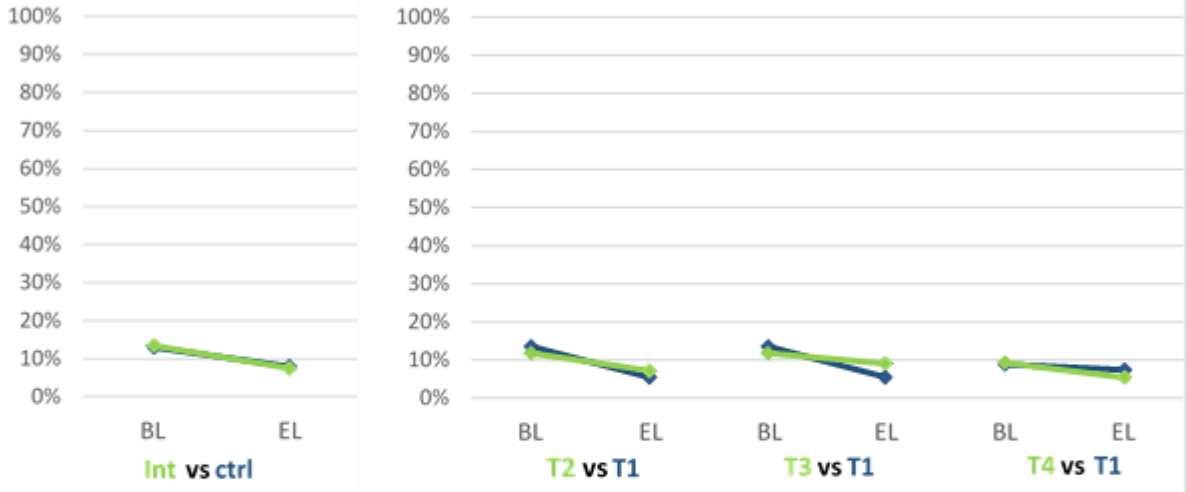
[14] http://www.who.int/childgrowth/standards/technical_report/en/ . Stata's zscore06 package was used to calculate scores.

Box 29: Proportion of children aged 0-59 months with severe acute malnutrition

Specification 1 (quasi-experiment): Intervention (pooled) vs control

Specification 3 (experiment): T2, T3, and T4 vs T1

BL n=9,413
EL n=12,440



All	-0.010
Poor	-0.066*

All	0.036	0.054**	0.011
Poor	-0.046	0.051	0.033

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	-0.033	0.004	0.022	0.010
Poor	-0.065**	-0.123*	-0.028	-0.037*

Stratified analysis on the poorest 20%. Among the poorest 20%, effect estimates for the impact of PBF compared to status quo are small positive overall as well as for intervention arms T1 and T2 (indicating small negative impact), but not statistically significant. No added benefit of the "adds-on" beyond the standard T1 is apparent for the poorest 20%.

Indicator 29: Impact of PBF on the proportion of children aged 0-59 months with severe acute malnutrition

Sample, indicator measurement and calculation. For this indicator, the sample was restricted to all children aged 0-59 months for whom weight and height were measured. Children were considered to suffer from severe acute malnutrition if their weight-for-height z-scores were below three standard deviations from the reference median based on the 2006 WHO child growth standards[14].

Main results. Results pertaining to indicator 29 are displayed in **Box 29**. The proportion of children with severe acute malnutrition decreased slightly from an 13% to 8% between baseline and endline in both the intervention and control groups. The corresponding impact estimates (specification 1 and 2) indicate no impact of PBF compared to status quo. The experimental study component (specification 3) indicates a small incremental effect of T3 over and above T1, in that malnutrition rates did not decrease quite as much in T3 as in T1.

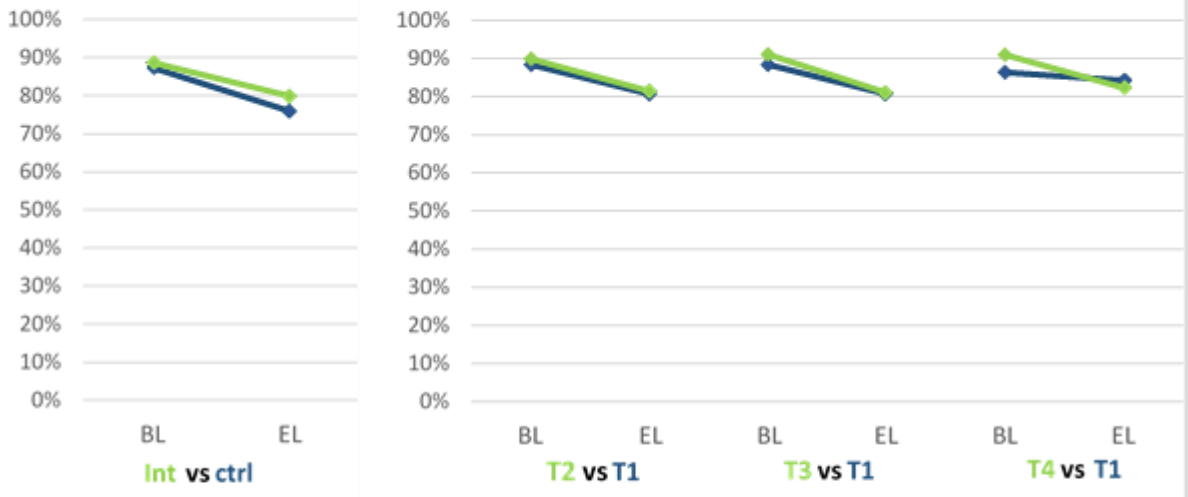
Stratified analysis on the poorest 20%. Contrary to the overall sample, we detected an effect for the comparison of PBF against status quo overall (-6.6 pp) and particular in intervention arms T1 (-6.5 pp) and T2 (-12.3 pp), in that PBF decreased malnutrition rates. No added benefit of the "adds-on" beyond the standard T1 is apparent for the poorest 20%.

Box 30: Proportion of children aged 6-59 months with anemia

Specification 1 (quasi-experiment): Intervention (pooled) vs control

Specification 3 (experiment): T2, T3, and T4 vs T1

BL n=8,788
EL n=11,281



All	0.034
Poor	-0.012

All	-0.012	-0.025	-0.022
Poor	-0.088*	-0.032	-0.076

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	0.033	0.039	0.026	0.042
Poor	0.001	-0.048	0.004	-0.019

Indicator 30: Impact of PBF on the proportion of children aged 6-59 months with anemia

Sample, indicator measurement and calculation. For this indicator, the sample was restricted to all children aged 6-59 months for whom an anemia test was performed. In accordance with WHO standards[15], children with hemoglobin levels below 11 g/dl were considered anemic.
Main results. Results pertaining to indicator 30 are displayed in **Box 30**. The proportion of anemic children decreased from an 88% to 79% between baseline and endline overall, slightly less so in the intervention group. The corresponding impact estimates (specification 1 and 2) are not statistically significant, however. No added benefit of the "adds-on" beyond the standard T1 (specification 3) could be detected in regards to anemia in children under 5.

Stratified analysis on the poorest 20%. In the subsample of the poorest 20%, impact estimates are also near zero, but on the negative side, meaning that anemia rates decreased slightly more in PBF areas. The direct comparison of the different intervention arms shows an incremental benefit of T3 over and above T1 (-8.8 pp).

Additional analyses. As robustness check, we further performed the analysis considering only children with moderate or severe anemia (hemoglobin levels below 10 g/dl) as anemic. Results are fully aligned with those for all levels of anemia.

Indicator 31: Impact of PBF on the proportion of women aged 15-49 years with moderate or severe anemia

Sample, indicator measurement and calculation. For this indicator, the sample was restricted to all women for whom an anemia test was performed. In accordance with WHO standards[15], women were considered anemic if they had hemoglobin levels below 12 g/dl if not pregnant, and below 11 g/dl if pregnant.

Main results. Results pertaining to indicator 31 are displayed in **Box 31**. The proportion of anemic women decreased from an 65% to 60% between baseline and endline overall, slightly less so in the intervention group. The corresponding impact estimates (specification 1 and 2) are not statistically significant, however. This is true for all intervention arms, with the exception of a significant effect estimate for T4, which is largely attributable to an overall lack of change in the Nouna and Solenzo districts. No added benefit of the "adds-on" beyond the standard T1 (specification 3) could be detected in regards to anemia in women.

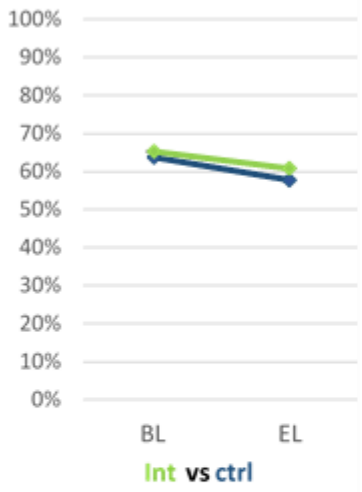
Stratified analysis on the poorest 20%. In the subsample of the poorest 20%, impact estimates indicate that anemia rates decreased slightly more in PBF areas, particularly in T2 and T3. None of the estimates are statistically significant, however.

Additional analyses. We further performed the analysis considering only women with moderate or severe anemia (hemoglobin levels below 11 g/dl (non-pregnant)/10 g/dl (pregnant)) as anemic. Results are fully aligned with those for all levels of anemia.

[15] http://www.who.int/vmnis/indicators/haemoglobin_fr.pdf

Box 31: Proportion of women aged 15-49 years with moderate or severe anemia

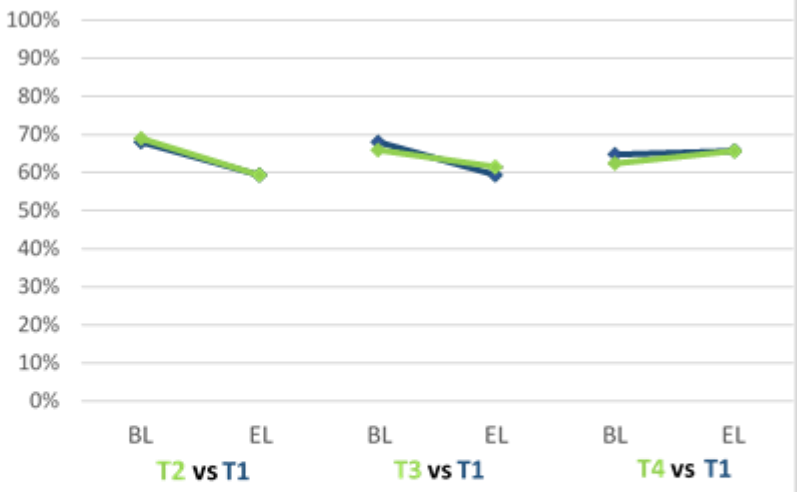
Specification 1 (quasi-experiment): Intervention (pooled) vs control



All	0.012
Poor	-0.029

Specification 3 (experiment): T2, T3, and T4 vs T1

BL n=7,975
EL n=10,195



All	-0.009	0.039	0.056
Poor	-0.052	-0.042	-0.129

Specification 2 (quasi-experiment): Intervention vs control

	T1 vs ctrl	T2 vs ctrl	T3 vs ctrl	T4 vs ctrl
All	0.016	-0.040	0.008	0.106**
Poor	0.011	-0.080	-0.069	-0.018

Summary: Impact of PBF on population health indicators

Table 16 and **Table 17** summarize impact estimates for the four population health indicators. Unlike before negative significant impact estimates are marked green, indicating an PBF-attributable decline in illness rates. Conversely, positive significant effect estimates signaling negative impact are marked in red. Cells not marked in color contain estimates that did not reach statistical significance.

Table 16: Summary of results pertaining to the impact of PBF on population health indicators (full sample)

	Quasi-experiment					Experiment		
	PBF vs control	T1 vs control	T2 vs control	T3 vs control	T4 vs control	T2 vs T1	T3 vs T1	T4 vs T1
28: Severe stunting (children U5)	-0.012	-0.001	-0.002	-0.021	-0.061***	0.041**	0.022	-0.018
29: Severe acute malnutrition (U5)	-0.010	-0.033	0.004	0.022	0.010	0.036	0.054**	0.011
30: Children U5 with anemia	0.034	0.033	0.039	0.026	0.042	-0.012	-0.025	-0.022
31: Women 15-49 with anemia	0.012	0.016	-0.040	0.008	0.106**	-0.009	0.039	0.056

Note: * p < 0.1, ** p < 0.05, *** p < 0.01. All effect estimates pertain to absolute (as opposed to relative) change. Effect estimates can be converted to percentages and reflect percentage point changes.

Table 17: Summary of results pertaining to the impact of PBF on population health indicators (poorest 20%)

	Quasi-experiment					Experiment		
	PBF vs control	T1 vs control	T2 vs control	T3 vs control	T4 vs control	T2 vs T1	T3 vs T1	T4 vs T1
28: Severe stunting (children U5)	0.030	0.052	0.061	-0.026	-0.006	0.023	-0.058	0.084
29: Severe acute malnutrition (U5)	-0.066*	-0.065**	-0.123*	-0.028	-0.037*	-0.046	0.051	0.033
30: Children U5 with anemia	-0.012	0.001	-0.048	0.004	-0.019	-0.088*	-0.032	-0.076
31: Women 15-49 with anemia	-0.029	0.011	-0.080	-0.069	-0.018	-0.052	-0.042	-0.129

Note: * p < 0.1, ** p < 0.05, *** p < 0.01. All effect estimates pertain to absolute (as opposed to relative) change. Effect estimates can be converted to percentages and reflect percentage point change

4. Discussion

Appraising the impact of the PBF program at once is no standard task, considering the multiple indicators included and the variety of effects observed across these indicators. Nevertheless, as we conclude this report, we try to draw a story line to bring together what can be learned from the Burkinabé PBF experience. In doing so, we look at general patterns across similar indicators, focusing primarily, but not exclusively on indicators for which we detected statistical significance (see 3.2). In this section, we purposely refer to existing literature only to a minimal extent when really needed, since our objective is to explain findings in the light of the country contextual elements related to the implementation of PBF in Burkina Faso. In particular, we feel the responsibility to discuss results in relation to the major user fee reduction policy (gratuité) introduced countrywide starting in June 2016. While being operative in the entire country means that the gratuité does not interfere with the identification of the effect attributable to PBF, it would be naïve to imagine that such a major health financing reform would not interact with the ongoing PBF pilot.

PBF appears to have produced a considerable impact on utilization of maternal care services, particularly delivery and PNC services, and on child and adult consultation (albeit not significant), but not really on the utilization of preventive child health services, where we observed no effect on vaccination and mixed effects on growth monitoring, positive in tendency for children under 1, negative for children ages 12-23 months (both not significant). Moreover, only for child and adult curative services, the experimental component of our study allowed us to detected a significant comparative advantage of the intervention arm combining PBF with insurance compared to the standard PBF intervention.

Before going into detail on the specific findings and potential explanations, we would like to underline again a number of methodological challenges which should be kept in mind when interpreting the results. First, comparisons of PBF against status quo are limited by the low number of clusters. This has implications both in that the minimum effect sizes the study is able to detect are relatively high, and in that there is a risk for imprecise estimation and faulty inference. We tested for the latter using the wild bootstrap technique. While all but one statistically significant effects appear to be robust, most confidence intervals contain zero and effects therefore need to be interpreted with some caution. Second, a fundamental prerequisite of inference of intervention impact is that each unit of observation is clearly in treatment or non-treatment, without any spillover. However, it appears that there was some spillover at regional level in that there was exchange among district health officers and competition between districts, leading to increased efforts even in control districts attributable to PBF. While this is certainly a desirable unintended effect of PBF, it might have led to an underestimation of the 'uncontaminated' intervention effects. In particular, it might have contributed to some of the apparent negative intervention effects driven by particularly strong positive change in control districts such as on quality of care for children. Further, unbiased inference of impact requires that no treatment similar to the intervention in question should have taken place at the same time. In a context in which not only the government continues to implement changes

to improve access and quality of care, but where a multitude of donors and non-governmental organizations are active, this was impossible to achieve. Not only was the gratuité policy implemented nation-wide in June 2016 as discussed in the introduction, but a variety of other interventions pertaining to reproductive and child health was on-going in both intervention and control districts in parallel to PBF. Effect estimates therefore likely do not only reflect the pure impact of PBF, but also at least to some extent the concurrent implementation of PBF with other interventions with similar objectives. This is in particular true for the effect estimates pertaining to the impact of PBF compared to status quo (specifications 1 and 2). Further qualitative research will be instrumental in understanding how spillover between districts and concurrent interventions might have influenced the results. Finally, when testing for impact on multiple indicators simultaneously, with an increasing number of indicators and impact estimates, there is an increasing risk of erroneously inferring intervention impact on some indicators due to the error margin accepted in each individual test ('multiple comparisons problem'). At particular risk of such erroneous inference are indicators with estimates of only marginal significance (only one *); these should therefore be interpreted with particular care. In further refining the analyses for public dissemination, the study team will attempt to rule out potential inference errors by applying state-of-the-art bootstrapping techniques [42].

The results must further be interpreted in light of implementation challenges experienced in the course of the intervention. Narrative evidence from implementers as well as the results of a parallel process evaluation led by the University of Montreal underline various implementation challenges which have likely hampered intervention effects. Such challenges include for instance substantial delays in payments resulting in frustration among healthcare personnel, delays and budgetary limitations in regards to the contract management and verification agents, and unintended dynamics introduced by the indigent selection process and community verification. Details can be found in [36] and [38]-[41].

Beyond these methodological and known implementation challenges, being able to explain the results is beyond the scope of a quantitative analysis and will represent the focus of our further qualitative work. However, as an initial working hypothesis, it is possible that the pattern of responses on service utilization we observe for PBF might have been influenced by the parallel implementation of the gratuité program at the national level. A potential emerging proposition, to be tested by qualitative research, is that healthcare providers focused on provision of services for which both the gratuité and the PBF program provided an explicit financial incentive, i.e. maternal care services and curative services for children and pregnant and lactating women. Both sets of services were subject to the payment of user fees till June 2016, when the Ministry of Health started to reimburse them on a fee-for-service basis, effectively introducing an additional incentive to provision on top of PBF payments. Preventive child health services, in contrast, have long been provided free of charge and as such were not subject to any additional payment once the gratuité was introduced in the country. In line with this observation postulating a potential interaction between PBF and gratuité, it is not surprising that we observed hardly any additional benefit of the intervention arms combining PBF with equity measures. These equity measures were

in fact removed for all services included in the gratuité, hence effectively equating the more complex PBF arms to the standard PBF in T1.

Another potential emerging hypothesis is that PBF could more effectively produce change on services which had long been the target of national policies, such as maternal care services which had been the target of the SONU (Soins obstétricaux et néonataux d'urgence) policy starting in 2007, possibly due to a certain readiness among healthcare providers to enable change. A similar argument could explain why the intervention arm combining PBF with insurance was able to produce greater changes in utilization of curative services than standard PBF. It is plausible that the mere presence of the insurance sensitized providers and communities on the importance of health service use, fully seizing the benefits of PBF once the gratuité was also rolled out.

This series of observations calls into question the value of assessing the impact of single interventions and points at the need to consider more broadly what happens at the health system level and how interventions working to change demand-side behaviors on a large scale (such as the gratuité) may allow to seize the full benefit of interventions working on the supply side to alter providers' behavior (such as PBF).

Of note is the fact that while PBF did not produce an effect on use of modern family planning at a general population level, possibly due to wider family planning programs, it was effective in ensuring higher update of modern family planning methods among the poorest quintile of the population. Again, this indicates the potential benefit of intervening with a specific supply-side intervention on top of national programs acting on the demand side.

PBF appears to have produced mixed impacts in terms of quality of service delivery. It is important to note, however, that the indicators chosen for the purpose of the impact evaluation are not fully aligned with the quality indicators incentivized by PBF. The latter largely contain indicators related to availability of inputs and process indicators based on document review. The impact evaluation, in contrast, relied on direct observation of actual care provided. For instance, PBF incentivizes correct use of the PCIME checklist as determined by document review, whereas our corresponding indicator pertains to adherence to PCIME as directly observed, irrespective of the checklist. While actual care provided ultimately is what the intervention aims at, it is important to remember this slight misalignment between what PBF purchased and the quality of care indicators used for the purpose of the impact evaluation. In general, the changes observed in the quality of service delivery are well below expectation with significant positive intervention effects being reported only for structural elements (water and electricity supply) and for completeness of routine ANC services. The lack of impact on drug availability may at first appear surprising, because one would expect the additional resources generated by PBF to be easily deployed towards improving drug availability. However, discussions with implementing stakeholders illustrated that this was rarely the case since the introduction of PBF was not accompanied by measures to enable providers to procure drugs outside the standard government supply chain. Interestingly, the positive impact observed on ANC routine care is not matched by women's reports, indicating a relative decline in perceived quality of care produced by stable satisfaction

levels in intervention facilities against increases in perceived quality of care in controls. This discrepancy might be the result of the fact that women's expectations on the potential of PBF to stimulate change were not met by the reality of the project implementation.

The negative impact on completeness of IMCI routine symptoms checking is particularly worrisome and we wonder to what extent it may be driven by an increased workload. We see from the results presented above that PBF did lead to an increase in utilization of curative services (albeit not reaching significance). It is possible that this increase in service utilization was of a larger magnitude than expected, due to the synergetic effect of the gratuité, bringing facilities to operate at their maximum capacity and hence hampering their ability to maintain quality standards. At the same time, beyond the effect estimates themselves, we know that the negative effect on quality of service delivery observed in intervention districts for child services is largely driven by substantial quality improvements in control districts, particularly in two districts, Barsalogho and Ziniaré. In Barsalogho, an intensive intervention to improve health care for children has been on-going since 2008. Although baseline quality levels were very poor, it is possible that PBF introduced the necessary incentives to effect translation of capacity building efforts in the context of the prior intervention into practice. Understanding what further interventions took place in these districts and how they might have produced quality improvements beyond those produced by PBF may be the object of further qualitative research.

PBF also appeared to produce hardly any effect on indicators pertaining to health status. The only consistent effect detected is a reduction in the proportion of children with severe acute malnutrition among the poorest quintile, suggesting that, similarly to what discussed earlier in relation to use of modern family planning methods, while PBF might have not stimulated changes at the population level, it has done so among the poorest. Interestingly and again in line with the effect detected on use of modern family planning methods, this pro-poor effect was consistent across intervention arms and not tied to the equity measures implemented in some selected intervention arms. This may again have to do with what already mentioned earlier, i.e. the fact that beyond the initial targeting effort which in and of itself represents a major intervention probably raising awareness on issues pertaining to equity in health, T1, T2, and T3 were not as highly differentiated in practice as in principle due to the manner in which the additional payments/incentives were actually applied. Note that we cannot exclude that the potential sampling bias discussed in 2.5 might have somewhat affected impact estimates on population health indicators. We therefore recommend interpretation with caution.

The lack of effect on indicators pertaining to health status calls into question the pertinence of even expecting that a program may produce such changes in such a limited time period. Albeit any intervention targeting the health system ultimately aims at improving population health, it is plausible to assume that given long pathways to changing health, measuring the impact of PBF on actual health outcomes may be beyond the scope of a three-year impact evaluation. This remark appears to find support in the very few PBF evaluations having focused on health impacts [13][26][27] [28]. In addition, we must consider that the disease burden associated with

malnutrition and anemia is tied to wider eco-social determinants of health, well beyond the role that improvements in provision of quality care can play.

The impact detected on dimensions related to human resources is probably also well below expectations, with satisfaction patterns improving consistently only in relation to the physical work environment, possibly due to improvements in infrastructure or availability of equipment and material made possible by the additional revenues generated through PBF. On other dimensions, we did not find any impact of PBF. To be noted is the fact that the lack of impact on intrinsic motivation can be interpreted as a positive feature of the program, since it indicates that introducing incentives attached to performance does not erode health workers' intrinsic work motivation as feared by some [29]. Research from other settings (e.g. [30]-[35]) suggests that the psychological and motivational mechanisms of PBF constitute a complex interplay of diverse positive and negative factors highly dependent on health workers' experiences of the specific intervention design, implementation, and implementation context. The descriptive results on health workers' perceptions of PBF presented in 3.1 suggest substantial variation in how health workers have experienced and evaluated the intervention overall and its various components. Although findings from a parallel process evaluation [36][37] have provided some insights, we currently lack a sufficiently profound understanding of how the intervention unfolded and was perceived by actor on the ground. A more in-depth exploration of operational factors and perceptions might therefore be an interesting area for further research in the quest to better understand how the PBF "black box" operates to effect changes in health service provision.

Beyond impact on the single dimensions, the results of the impact evaluation highlight two striking features: the limited additional benefit of combining PBF with additional equity measures as in T2 and T3, and the somewhat different change in T4, combining PBF with health insurance, compared to the other intervention arms.

On the one side, the motives explaining the limited additional benefit of combining PBF with additional equity measures have been discussed above. Primarily, we postulate that the three intervention arms (T1, T2, and T3) differed from one another much later than what originally expected, since additional payments/rewards were in practice only attached to a limited set of indicators. Although a publication on the unintended effects of targeting has already been published [38], the qualitative study planned to complement the impact evaluation will explore issues related to targeting and to the complexity of implementing three parallel arms at once. There is a clear need to understand to what extent the lack of differentiation across the three arms are linked to design issues or to implementation challenges.

On the other side, repeatedly throughout the evaluation, it appears that T4 facilities displayed a somewhat different pattern as all other intervention arms compared to controls, for instance in regards to intrinsic motivation, ANC patient education, or ANC utilization. Often, results from the experimental component of the study ("T4 vs T1" comparison) seem somewhat contrary in that the apparent superiority or inferiority of T4 is not reflected there. As pointed out in the description of results, however, we believe that issues rooted in the study design limit the interpretability of the "T4 vs control" estimates. Although not discussed explicitly, data show substantial variation

in change over time between districts on most indicators (see **Appendix F**). In specification 2 models, facilities and catchment areas from the different intervention arms are compared to controls from all 12 control districts. Assuming that control districts are generally and on average good counterfactuals, the comparison of T1 against controls is not affected by between-district variation in change, as this variation is averaged out across the 12 districts in which T1 was implemented. Similarly, between-district variation is averaged out across the 8 districts in which T2 and T3 was implemented. However, T4 was implemented only in two districts, Nouna and Solenzo in the Boucle du Mouhoun region. Although these two districts overall do not appear to be generally exceptional, they have experienced substantial above- or below-average change on a number of indicators when compared to other intervention districts, in most cases uniformly across their T1 and their T4 facilities and catchment areas. Such cases have little influence on the "T1 vs controls" comparison (specification 2), since Nouna and Solenzo are only two of 12 districts in which T1 is implemented. However, implications are stronger for the "T4 vs controls" comparison, where all T4 facilities come from Nouna and Solenzo, whereas controls come from all 12 control districts. Specification 2 model estimates then suggest particular impact of T4, which however does not appear to be due to T4 as such, but rather to particular secular trends in the two districts, against which the controls do not appear to be optimal counterfactuals. We therefore urge for caution in interpreting the "T4 vs control" estimates. As discussed in the introduction, from a methodological point of view, a randomization of T4 across all/most intervention districts as done for T2 and T3 would have been preferable, but was not possible for feasibility reasons. One possible solution would be an additional analysis limiting the "T4 vs controls" comparison only to the Boucle du Mouhoun region or selecting appropriate controls by other means. However, we have decided against such additional analyses as the number of clusters would be very small and statistical power very low, accordingly.

Looking only at results regarding the additional value of T4 compared to the standard T1, as tested experimentally in the Nouna and Solenzo districts, we found positive impact only in regards to tetanus vaccinations in pregnancy and curative consultations, and negative impact only in regards to perceived quality of care of consultations of children under 5 and PNC utilization. Explanation of both impact on those indicators and lack thereof on the others is difficult and will require additional qualitative research. One likely explanation for the relative lack of added value of T4 compared to T1 might be that actual insurance enrolment rates remained extremely low throughout the course of the intervention. Again, the "T4 vs T1" findings need to be interpreted in knowledge of that fact that not all T4 facilities in Nouna and Solenzo were randomized and therefore entered the analysis. It is possible that prior experience with insurance facilitated implementation. Sample sizes are too small for robust quantitative analyses in this regard, but this is an interesting question for further qualitative exploration.

In conclusion, our impact evaluation revealed that PBF resulted in mixed effects compared to status quo, producing positive change on some indicators and not on others. In addition, our impact evaluation identified little added value of the equity measures which were combined with PBF in some intervention arms. In line with what outlined in the discussion section, further qualitative research is necessary to

gain a better understanding of the patterns detected by the impact evaluation and formulate policy recommendations accordingly.

5. References

[1] Su, T., Kouyaté, B., & Flessa, S. (2006). Catastrophic household expenditure for health care in low-income society: a study from Nouna District, Burkina Faso. *Bulletin Of The World Health Organisation*, *84*(1), 21-27. doi: 10.2471/blt.05.023739

[2] Beogo, I., Huang, N., Gagnon, M., & Amendah, D. (2016). Out-of-pocket expenditure and its determinants in the context of private healthcare sector expansion in sub-Saharan Africa urban cities: evidence from household survey in Ouagadougou, Burkina Faso. *BMC Research Notes*, *9*(1). doi: 10.1186/s13104-016-1846-4

[3] Dong, H., Gbangou, A., De Allegri, M., Pokhrel, S., & Sauerborn, R. (2008). The differences in characteristics between health-care users and non-users: implication for introducing community-based health insurance in Burkina Faso. *The European Journal Of Health Economics*, *9*(1), 41-50. doi: 10.1007/s10198-006-0031-4

[4] Pokhrel, S., De Allegri, M., Gbangou, A., & Sauerborn, R. (2010). Illness reporting and demand for medical care in rural Burkina Faso. *Social Science & Medicine*, *70*(11), 1693-1700. doi: 10.1016/j.socscimed.2010.02.002

[5] De Allegri, M., Tiendrebéogo, J., Müller, O., Yé, M., Jahn, A., & Ridde, V. (2015). Understanding home delivery in a context of user fee reduction: a cross-sectional mixed methods study in rural Burkina Faso. *BMC Pregnancy And Childbirth*, *15*(1). doi: 10.1186/s12884-015-0764-0

[6] Atchessi, N., Ridde, V., & Zunzunegui, M. (2016). User fees exemptions alone are not enough to increase indigent use of healthcare services. *Health Policy And Planning*, *31*(5), 674-681. doi: 10.1093/heapol/czv135

[7] Kadio, K., Ridde, V., & Mallé, S. (2014). [The difficulties of access to health care for indigent living in non-poor households]. *Santé Publique*, *26*(1), 89-97. doi: 10.3917/spub.137.0089

[8] Steenland, M., Robyn, P., Compaore, P., Kabore, M., Tapsoba, B., & Zongo, A. et al. (2017). Performance-based financing to increase utilization of maternal health services: Evidence from Burkina Faso. *SSM Population Health*, *3*, 179-184. doi: 10.1016/j.ssmph.2017.01.001

[9] Ministère de la Santé Burkina Faso. (2013). Guide de mise en oeuvre du financement base sur les résultats dans le secteur de la santé. Retrieved from http://www.fbrburkina.org/articles/item/6.html

[10] Ridde, V., Yaogo, M., Kafando, Y., Sanfo, O., Coulibaly, N., Nitiema, P., & Bicaba, A. (2009). A community-based targeting approach to exempt the worst-off from user fees in Burkina Faso. *Journal Of Epidemiology & Community Health*, *64*, 10-15. doi: 10.1136/jech.2008.086793

[11] Beaugé, Y., Koulidiati, J., Ridde, V., Robyn, P., & De Allegri, M. How much does community-based targeting of the ultra-poor in the health sector cost? Novel evidence from Burkina Faso. Under review.

[12] Institute of Public Health, Heidelberg University. (2016). *Final Results of the RBF4MNH Impact Evaluation.* Retrieved from https://www.klinikum.uni-heidelberg.de/fileadmin/inst_public_health/Dokumente/Final_Results_Report.pdf

[13] The World Bank. (2017). *Cameroon Performance-Based Financing Impact Evaluation Report*. Retrieved from http://documents.worldbank.org/curated/en/756781499432127674/pdf/117301-WP-P126389-PUBLIC-Cameroon.pdf

[14] Institute of Public Health, Heidelberg University. (2015). *Impact Evaluation for Health Performance-Based Financing in Burkina Faso. Baseline report.* Retrieved from http://microdata.worldbank.org/index.php/catalog/2761/related_materials

[15] Angrist, J., & Pischke, J. (2009). *Mostly harmless econometrics*. Princeton, N.J: Princeton University Press.

[16] Ouédraogo, S., Ridde, V., Atchessi, N., Souares, A., Koulidiati, J., Stoeffler, Q., & Zunzunegui, M. (2017). Characterisation of the rural indigent population in Burkina Faso: a screening tool for setting priority healthcare services in sub-Saharan Africa. *BMJ Open*, *7*(10), 1-9. doi: 10.1136/bmjopen-2016-013405

[17] Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How Much Should We Trust Differences-In-Differences Estimates? *The Quarterly Journal Of Economics*, *119*(1), 249-275. doi: 10.1162/003355304772839588

[18] Cameron, A., Gelbach, J., & Miller, D. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *Review Of Economics And Statistics*, *90*(3), 414-427. doi: 10.1162/rest.90.3.414

[19] Carter, A., Schnepel, K., & Steigerwald, D. (2017). Asymptotic Behavior of a t-Test Robust to Cluster Heterogeneity. *The Review Of Economics And Statistics*, *99*(4), 698-709. doi: 10.1162/rest_a_00639

[20] Imbens, G., & Kolesár, M. (2016). Robust Standard Errors in Small Samples: Some Practical Advice. *Review Of Economics And Statistics*, *98*(4), 701-712. doi: 10.1162/rest_a_00552

[21] Colin Cameron, A., & Miller, D. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal Of Human Resources*, *50*(2), 317-372. doi: 10.3368/jhr.50.2.317

[22] Angrist, J., & Lavy, V. (2002). The effect of high school matriculation awards: Evidence from Randomized Trials (NBER Working Paper 9389). Cambridge, MA: National Bureau of Economic Research.

[23] Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal Of Statistics*, *9*(2), 139-158. doi: 10.2307/3314608

[24] Lohmann, J., Souares, A., Tiendrebéogo, J., Houlfort, N., Robyn, P., Somda, S., & De Allegri, M. (2017). Measuring health workers' motivation composition: Validation of a scale based on Self-Determination Theory in Burkina Faso. *Human Resources for Health, 15*, 33. doi: 10.1186/s12960-017-0208-1

[25] Donabedian, A. (1988). The quality of care. How can it be assessed? The Journal Of The American Medical Association, 260(12), 1743-1748. doi: 10.1001/jama.260.12.1743

[26] Lim, S. S., Dandona, L., Hoisington, J.A., James, S.L., Hogan, M.C., & Gakidou, E. (2010). India's Janani Suraksha Yojana, a conditional cash transfer programme to increase births in health facilities. An impact evaluation. *The Lancet, 375,* 2009–23. doi: 10.1016/S0140-6736(10)60744-1

[27] Powell-Jackson, T., Neupane, B. D., Tiwari, S., Tumbahangphe, K., Manandhar, D., & Costello, A.M. (2009). The impact of Nepal's national incentive programme to promote safe delivery in the district of Makwanpur. *Advances in Health Economics and Health Services Research, 21,* 221–49.

[28] Bowser, D., Gupta, J., & Nandakumar, A. (2016). The effect of demand- and supply-side health financing on infant, child, and maternal mortality in low- and middle-income countries. *Health Systems & Reform, 2,* 147–59. doi: 10.1080/23288604.2016.1166306

[29] Ireland, M., Paul, E., & Dujardin, B. (2011). Can performance-based financing be used to reform health systems in developing countries? *Bulletin Of The World Health Organization*, *89*(9), 695-698. doi: 10.2471/blt.11.87379

[30] Lohmann, J., Houlfort, N., & De Allegri, M. (2016). Crowding out or no crowding out? A Self-Determination Theory approach to health worker motivation in performance-based financing. *Social Science & Medicine*, *169*, 1-8. doi: 10.1016/j.socscimed.2016.09.006

[31] Lohmann, J., Wilhelm, D., Kambala, C., Brenner, S., Muula, A., & De Allegri, M. (2017). "The money can be a motivator, to me a little, but mostly PBF just helps me to do better in my job." An exploration of the motivational mechanisms of performance-based financing for health workers in Malawi. *Health Policy And Planning*, *33*(2), 183-191. doi: 10.1093/heapol/czx156

[32] Bhatnagar, A., & George, A. (2016). Motivating health workers up to a limit: partial effects of performance-based financing on working environments in Nigeria. *Health Policy And Planning*, *31*(7), 868-877. doi: 10.1093/heapol/czw002

[33] Paul, E., Sossouhounto, N., & Eclou, D. (2014). Local stakeholders' perceptions about the introduction of performance-based financing in Benin: a case study in two health districts. *International Journal Of Health Policy And Management*, *3*(4), 207-214. doi: 10.15171/ijhpm.2014.93

[34] Lagarde, M., Burn, S., Lawin, L., Bello, K., Dossou, J.-P., Makoutode, P., Goufodji, B. S., Lemière, C., & Juqouis, M. (2015). *Exploring the impact of*

performance-based financing on health workers' performance in Benin. Retrieved from https://www.rbfhealth.org/sites/rbf/files/Benin%20RBFHRH%20report.pdf

[35] Shen, G.C., Nguyen, H.T.H., Das, A., Sachingongu, N., Chansa, C., Qamruddin, J., & Friedman, J. (2017). Incentives to change: effects of performance-based financing on health workers in Zambia. *Human Resources for Health, 15,* 20. doi: 10.1186/s12960-017-0179-2

[36] Ridde, V., Yaogo, M., Zongo, S., Somé, P., & Turcotte-Tremblay, A. (2018). Twelve months of implementation of health care performance-based financing in Burkina Faso: A qualitative multiple case study. *The International Journal Of Health Planning And Management*, *33*(1), e153-e167. doi: 10.1002/hpm.2439

[37] Ridde, V., Turcotte-Tremblay, A., Souares, A., Lohmann, J., Zombré, D., & Koulidiati, J. et al. (2014). Protocol for the process evaluation of interventions combining performance-based financing with health equity in Burkina Faso. *Implementation Science*, *9*, 149. doi: 10.1186/s13012-014-0149-1

[38] Turcotte-Tremblay, A.-M., De Allegri, M., Gali-Gali, I. A., & Ridde, V. (2018). The Unintended Consequences of Combining Equity Measures with Performance-Based Financing in Burkina Faso. Accepted for publication in the International Journal of Equity in Health

[39] Bodson, O., Barro, A., Tucotte-Tremblay, A.-M., Zanté, N., Somé, P.-A., & Ridde, V. (2018). A study on the implementation fidelity of the performance-based financing policy in Burkina Faso after 12 months. *Archives of Public Health, 76:*4.

[40] Turcotte-Tremblay, A.-M., Gali-Gali, I. A., De Allegri, M., & Ridde, V. (2017). The unintended consequences of community verifications for performance-based financing in Burkina Faso. *Social Science & Medicine, 191,* 226-236.

[41] Fillol, A., Lohmann, J., Turcotte-Tremblay, A.-M., Somé, P.-A., & Ridde, V. Health workers' leadership and motivation in a results-based financing context in Burkina Faso: A multiple-case study. Under review in the International Journal of Health Policy and Management

[42] List, J. A., Shaikh, A. M., & Yang, X. (2016). Multiple Hypothesis Testing in Experimental Economics. home.uchicago.edu/amshaikh/webfiles/experimental.pdf. Accessed Jun 14 2018