CSEF

Centre for Studies in Economics and Finance

WORKING PAPER NO. 699

Assessing the Efficacy of a Results-Based Financing Project Targeted at the Pediatric Wards of Two Ugandan Hospitals

Sergio Beraldo, Michela Collaro, Elisabetta D'Agostino, Luigi Greco

Venice Omona, and Domenico Suppa

January 2024

University of Naples Federico II

University of Salerno

Bocconi University, Milan

CSEF - Centre for Studies in Economics and Finance
DEPARTMENT OF ECONOMICS AND STATISTICS – UNIVERSITY OF NAPLES FEDERICO II
80126 NAPLES - ITALY
Tel. and fax +39 081 675372 – e-mail: csef@unina.it
ISSN: 2240-9696

CSEF
Centre for Studies in Economics and Finance

WORKING PAPER NO. 699

Assessing the Efficacy of a Results-Based Financing Project Targeted at the Pediatric Wards of Two Ugandan Hospitals

Sergio Beraldo[*], **Michela Collaro**[†], **Elisabetta D'Agostino**[‡], **Luigi Greco**[§]
Venice Omona[**], **and Domenico Suppa**[††]

Acknowledgments: the authors are grateful to the Italian Agency for Development Cooperation (AICS), the Corti Foundation, the Ambrosoli Foundation and the Gulu University.

[*] University of Naples Federico II and CSEF. Email: s.beraldo@unina.it

[†] University of Naples Federico II. Email: michela.collaro@unina.it *(Corresponding author)*

[‡] Corti Foundation and St. Mary's Hospital Lacor. Email: e.dagostino@fondazionecorti.it

[§] University of Naples Federico II. Email: ydongre@unina.it

[**] St. Mary's Hospital Lacor. Email: veniceomon@gmail.com

[††] University Of Campania L. Vanvitelli. Email: domenicosuppa@gmail.com

1. Introduction

Achieving universal health coverage is one of the Sustainable Development Goals adopted by the United Nations in 2015 as part of the 2030 Agenda for Sustainable Development. The essence of universal health coverage is to grant access to health services to anyone, so avoiding that people with health care needs are exposed to the risk of financial ruin or impoverishment. Although the interventions aimed at reaching this ambitious goal by 2030 are everywhere problematic, in developing countries they appear even more so, for the health systems suffer from a severe lack of resources, reflected in shortages of adequate technology and specialized workforce (Di Pietro et al., 2020).

In this light, some attention has been recently devoted to the assessment of managerial models which may improve the delivery of services in contexts characterized by shortages of health care resources (Eldridge and TeKolste, 2016; Honda, 2013; United Nations, 2011). Among these models, particular interest has been placed on Results-Based Financing schemes (hereafter, RBF); these are mechanisms linking the provision of resources to the fulfilment of agreed upon performance standards.

In recent years, several developing countries have implemented RBF programs to enhance healthcare quality and increase the utilization of health services. However, uncertainties persist regarding their impact (e.g., Fretheim et al. 2012). These uncertainties primarily revolve around potential side effects associated with the use of financial incentives (Eldridge and TeKolste, 2016; Lemière et al., 2013; Oxman and Fretheim, 2008).

Firstly, connecting financial rewards to performance may incentivize the overproduction of rewarded tasks (e.g., fee-for-service payments can lead to unnecessary diagnostic tests) and the underproduction of equally vital tasks that are not directly remunerated. Secondly, recipients of bonuses may concentrate on populations that are easier to reach, leading to potential "cherry-picking"

of patients. Thirdly, performance incentives could exacerbate resource disparities both between and within countries, endangering healthcare access for the most vulnerable.

In 2018 the Corti Foundation, in partnership with the Ambrosoli Foundation, the St. Mary's Hospital Lacor, the Dr Ambrosoli Memorial Hospital Kalongo (hereafter Lacor and Kalongo Hospitals), the University of Naples and the University of Gulu, implemented a Results-Based Financing scheme in the Acholi region, Northern Uganda. This three-year project – funded by the Italian Agency for Cooperation and Development (AICS) – was targeted at the paediatric wards of the two aforementioned Private-Not-For-Profit health facilities (i.e. Lacor and Kalongo Hospitals) with the aim of enhancing the quality of inpatient healthcare services provided.

This paper seeks to evaluate whether the improvements in healthcare quality, attributed to the project (Greco et al., 2021), are indeed a result of the RBF initiative, and provide first evidence about the persistence of the effects related to the implementation of RBF schemes.

Differently from similar analyses, we rely on data concerning the quality of clinical and nursing procedures coming from a randomly extracted sample of clinical records (or charts, for short).

Evaluating the effectiveness of RBF programs typically involves examining performance indices (e.g., the number of visits or immunization coverage rates) or health indicators (e.g., maternal and child health, as seen in Mushasha and El Bcheraoui, 2023). These indicators are constructed based on information collected during periodic assessments that determine whether the recipient has met the quality standards set by the donor. However, one drawback of this approach is that the RBF mechanism may incentivize over-reporting of positive results.

In cases of highly disadvantaged situations, appointed commissioners assessing outcomes may be more lenient because a negative report would entail a loss of benefits, potentially resulting in restricted healthcare access for those in need. Therefore, it is crucial to employ evidence from alternative sources to assess performance accurately.

In this paper we rely on data from a randomly extracted sample of 1148 clinical records, filled in the 2014-2016 period (i.e., before the RBF project took place) and in 2020, the last year of the RBF project. Data coming from clinical records offer a wider range of details concerning the quality of clinical and nursing procedures delivered during the overall length of hospital stay. As an example, it is possible to evaluate, for each child admitted, the adequacy of the diagnosis procedures as well as the clinical tests carried out to detect hidden health(care) needs. Such details remain unknown when one considers the overall performance of the health unit and measures it by synthetic indicators observed at specific points in time.

The evidence presented in this paper indicates that, on the whole, the program enhanced both clinical and nursing procedures. Moreover, it underscores the significance of the timing of hospitalization. For instance, children admitted within 7 or 14 days prior to the periodic assessment of performance standards (commonly referred to as inspection days) are about five times as likely to belong to the group benefiting from the highest standard of clinical procedures. This improvement primarily stems from the enhanced appropriateness of treatments administered, rather than the appropriateness of diagnoses.

Similar results are observed when considering all hospitalizations that occurred 7 or 14 days after the inspections. In contrast, nursing procedures appear to be less influenced by periodic assessments.

We also document that the positive effects diminish in moving further away from the time of the periodic assessment, raising doubts regarding the persistence of the consequences brought about by RBF schemes.

The paper is organized as follows. In section 2, we briefly survey some key characteristics of the Results-Based Financing approach and the AICS pediatric RBF project in Northern Uganda. In Section 3, we introduce the data and the empirical strategy. In Section 4, we illustrate our findings. Section 5 discusses and concludes.

2. Result-Based Financing

2.1. Result-based financing: pros and cons

Results-based financing is a machinery that links payments (or material rewards) to the achievement of pre-defined performance standards. An RBF program can address either the demand side (individuals needing health care) through *conditional cash transfers*, or the supply side (healthcare providers or governments) through *performance-based aid*, *performance based-transfer*, *performance based contracts*.

Demand-side programs aim at improving access *to* and utilization *of* specific healthcare services or promote preventive health behaviour. Supply-side schemes aim at increasing healthcare coverage or at improving health outcomes by delivering specific health services, possibly in rural and underserved regions.

Because transfers are conditional on results, supply-side RBF programs are commonly supervised to check whether the stated objectives are indeed achieved.

Clearly, the relationship between the donor (i.e., the financing institution) and the recipient of funding suffers from the typical distortions characterizing any principal-agent relationship. As it is well known, whenever one actor (principal) delegates a task to another (agent), the objectives of the two parties - and the information available to them - might differ substantially, affecting the efficient achievement of the principal's aims (e.g., Laffont & Martimort, 2009).

In this light, the RBF might be a powerful tool to increase the quality of healthcare whenever the incentive scheme adopted by the principal is able to align her objectives with the ones of the recipient (Savedoff & Partner, 2010). In order for this to come true, besides being properly designed, incentives must represent a substantial proportion of both patients' and workers' income, a requirement that is more easily met in low/middle-income countries (Oxman & Fretheim, 2008).

Grittner (2013) reviewed the empirical evidence concerning the implementation of RBF schemes in several low/middle-income countries (LMICs), documenting the efficacy of the approach in improving healthcare delivery and reducing health spending for the poor.

Shroff et al. (2017) identified the primary factors, ranging from design to implementation, that could either facilitate or impede the spread of Results-Based Financing (RBF) schemes. The authors advise exercising caution when identifying the key actors and stress the importance of ensuring that the program aligns with the institutional context. They also emphasize the need to strike a balance between an optimally designed program and its actual financial sustainability (see also, for example, Fretheim et al., 2012; Shen et al., 2017).

Behavioral economics has long challenged the perspective rooted in the acceptance of the fundamental *law of behavior*, which posits that higher material incentives result in greater effort and improved performance. Behavioral responses to material incentives are not the sort of mechanical reactions one might anticipate if assume that human behavior strictly adheres to its simplified *homo oeconomicus* counterpart. There are other factors, at least as influential as material incentives, which shape behavior, such as the desire to conform to socially accepted norms.

The provision of material incentives may, in many instances, prove counterproductive, because extrinsic rewards can displace intrinsic motivations, which are equally vital for eliciting the desired behavior. Material incentives can prove ineffective or even produce unintended negative consequences (e.g., Gneezy et al., 2011; Bowles, 2008; Titmuss, 1970).

For all these reasons, the effects of an RBF program cannot be taken for granted; the evaluation stage is thus essential to understand whether the program delivered what it promised or delivered something even contrary to expectations[2].

[2] Beside monetary incentives, many authors outline that the reaction of workers to 'results-based' schemes strongly depends on monitoring and supervision (e.g. Lemière et al., 2013). These might induce compliance with good practices because of the effect that the perception of being observed exerts on individuals (Chen et al., 2015). Supervision might

2.2. The RBF projects in Uganda

Uganda is a low-income country, ranked among the 25 poorest in the world (e.g. Harrington 2018). The Ugandan National Health System comprises 6.937 health facilities, owned in part by the Government (45.16%) and in part by private organizations (54.84%). The private sector encompasses both Private-For-Profit and Private-Not-For-Profit health providers. Healthcare financing relies on different sources: central and/or local government funds, private out-of-pocket expenses, donations and support programs.

Since 1990, the Ugandan health system has experienced several reforms to improve health outcomes and access to care. Although considerable improvements have been observed in many health (and healthcare) indicators, some are still at an unacceptable level, and a significant gap across regions persists. As an example, albeit under-five mortality remarkably declined from 137 deaths per 1000 live births in 2006 to 45,8 in 2019 (Mejía-Guevara et al., 2019), it is still higher than the Sustainable Development Goal's threshold (25 per 1000 live birth).

Since 2003, the Ugandan health system has benefitted from several RBF projects targeted at both the supply and the demand side of the healthcare market. Some prominent donors have conditioned their financial support to the adoption of this approach. Also the Ugandan Ministry of Health relies on RBF schemes to increase healthcare utilization and improve its quality.

Ssengooba et al. (2015) conducted a comprehensive review of the primary Results-Based Financing (RBF) programs carried out in Uganda from 2003 to 2015. The supply-side initiatives were aimed at Private Not-For-Profit healthcare facilities, while the demand-side projects centered on maternal and child healthcare. Overall, in the districts where these programs were implemented, both types of schemes produced favorable outcomes: supply-side projects, including the *Cordaid Project* and

also be a tool for professional development whenever it ameliorates job motivation and satisfaction (Rowe et al., 2005). Workers' reputation might also play a prominent role to improve procedures when individuals' performances become publicly known.

NuHealth, enhanced health service delivery, and demand-side projects contributed to a reduction in mortality rates.

However, these positive results were accompanied by noteworthy side effects. In general, RBF schemes targeting health facilities faced sustainability concerns, while those concentrating on healthcare users led to an unsustainable surge in healthcare service utilization, significantly increasing the demand. After a long pause following the NuHealth experience, in 2018 the Corti and the Ambrosoli Foundations, in partnership with the Lacor and Kalongo Hospitals, the University of Naples and the University of Gulu, implemented an RBF project in the Acholi region, Northern Uganda.

The project AID 11495, '*Result Based Financing, an engine of change for pediatric services*', was funded by the Italian Agency for Cooperation and Development (AICS) and lasted three years. The program was targeted at the Lacor and Kalongo hospitals' pediatric wards in order to improve the quality of inpatient healthcare provided.

The RBF scheme was designed in such a way as to condition financial benefits to performance, assessed on the basis of the value taken by well specified indices. These pertained to the number of children admitted monthly and to the assessment – made by a committee of experts - of five key dimensions related to healthcare quality: *infrastructure and organization*; *hygiene and cleanliness*; *clinical and nursing processes*; *emergency readiness*; *students training*.

Every three months, the committee of experts - three members of the Hospital Quality Assurance Department and a representative member of the Ugandan Ministry of Health – organized an *inspection day* at the children's wards in order to assess the relevant quality items.

Funds allocated through the RBF program were divided into two components: a basic share and a bonus. The basic share was contingent upon the number of hospital admissions, with a fixed amount allocated per child admitted. To calculate the bonus, the number of hospital admissions was

multiplied by a coefficient tied to the overall score obtained on quality indicators. The basic funds were utilized to cover general hospital expenses, while the bonus was partially designated to incentivize and reward hospital staff.

In particular, as far as the Kalongo (Lacor) hospital is concerned, 35% (25%) of the bonus share was assigned to the pediatric department staff and to the staff of the ancillary services to pediatric care, such as laboratory, radiology, etc. The remaining 65% (75%) was used to address shortcomings highlighted during the quarterly assessments.

All members of the pediatric staff received the same amount (1st category staff), regardless of their qualifications, duties and seniority, while members of the staff providing ancillary services to pediatrics (2nd category staff) received a lower amount; again: regardless of their qualifications, duties and seniority.

Remarkably, during the final year of the project, hospital managers decided to downgrade to the 2nd category those 1st category staff members who received a negative evaluation from medical direction or whose lack of commitment determined a low score in one of the items measured during the quality assessment.

According to the report released at the end of the project (Greco et al., 2021) – based on the evaluations made by the committee of experts at the quarterly evaluations - both hospitals exhibited an increase in overall performance because of the RBF implementation, so suggesting that the Result-Based schemes were effective in boosting the quality of care.

3. Data and methods

3.1. Descriptive evidence

The study presented in this paper relies on data from a sample of randomly extracted clinical charts. Specifically, we utilize information gathered from clinical charts that were completed during the

periods of 2014-2016 (prior to the commencement of the RBF project) and in 2020, which marks the final year of the RBF project.

The clinical charts included in the analysis were randomly extracted from the pool of available charts. Subsequently, an impartial evaluator assessed each chart, assigning a score to each item within a set of quality criteria associated with the appropriateness of clinical and nursing procedures.

The evaluator conducted this assessment without knowledge of the year in which the chart was completed. In terms of the number of charts analyzed, our study is based on 599 records to evaluate the appropriateness of clinical procedures (comprising 378 records from the 2014-2016 period and 221 records from the year 2020). Additionally, we examined 547 records to assess the appropriateness of nursing procedures (consisting of 341 records from the 2014-2016 period and 208 records from the year 2020). Anonymized copies of the original clinical charts used in this analysis can be made available upon request.

According to the score assignment rule reported in Table 1, the independent evaluator first assessed 11 items pertaining to the quality of clinical management (599 records were scrutinized in terms of both diagnoses and therapies). Then, using a different pool of 547 charts, the evaluator assessed 5 items concerning nursing care (the complete list of quality criteria can be found in Table A1 in the Appendix).

Table 1: Score assignment rule

-1	if absent, not done or not done according to guidelines
0	if missing or not applicable
1	if present or done, but unclear
3	if present or done, according to guidelines

Notes: *The total score associated with each medical chart is obtained by rescaling all the scores in a 0-3 interval and summing them.*

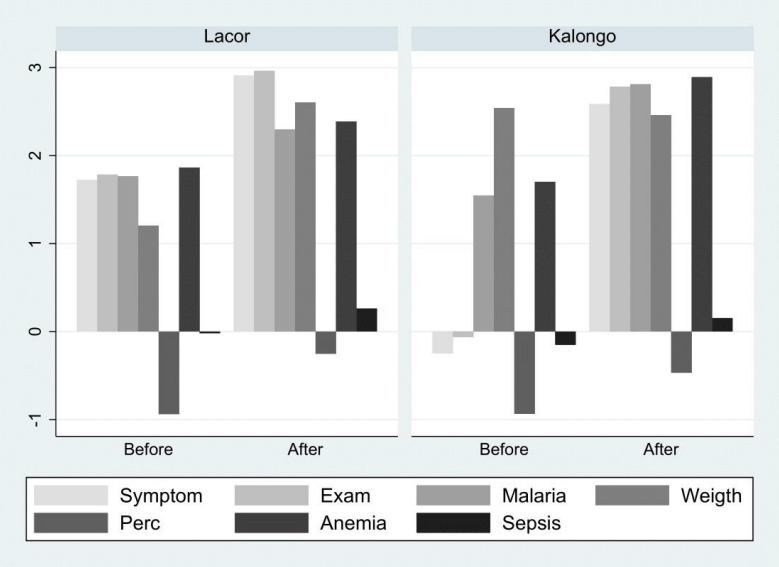
As far as our analysis is concerned, we first make a pre-post comparison, confronting the situation existing before the RBF program was implemented (2014-2016), with the one emerging in the final

year of the program (2020). This comparison explores potential differences between average scores in the quality items.

Figures 1 and 2 show the pre-post comparisons in terms of appropriateness of diagnoses and therapies, respectively. According to the data (see Figure 1), on average, diagnosis procedures improved in 2020 with respect to the pre-RBF situation. The account about the clinical history (*Symptom*) and the accurate examination of the children (*Exam*) remarkably improved at the Kalongo hospital, which started from a negative average score. A slight improvement is also evident in both hospitals as far as the diagnosis procedures for *Malaria*, *Anemia* and *Sepsis* are concerned.

As far as the adequacy of checking malnutrition is considered, although the score related to weight measurement (*Weight*) increased at Lacor hospital, no improvements were observed at Kalongo. The score assessing the provision of information about the weight centile (*Perc*) remained in the range [-1;0] - i.e., not available or inadequate data - at both hospitals, perhaps due to the lack of specific space on the clinical charts.

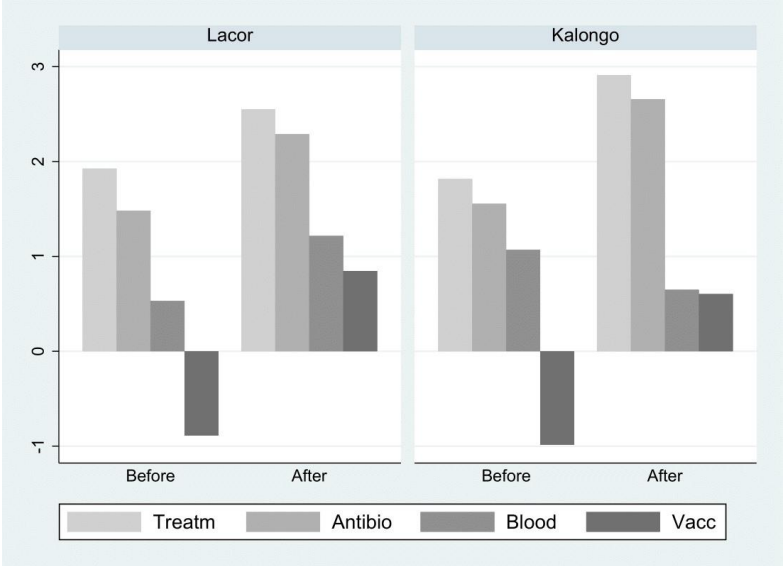
Figure 1. Pre-post comparison of diagnosis items' scores



Notes: *The figure on the left (right) displays the pre-post comparison of the average score attached to each quality item pertaining to diagnosis procedures at Lacor (Kalongo) hospital.*

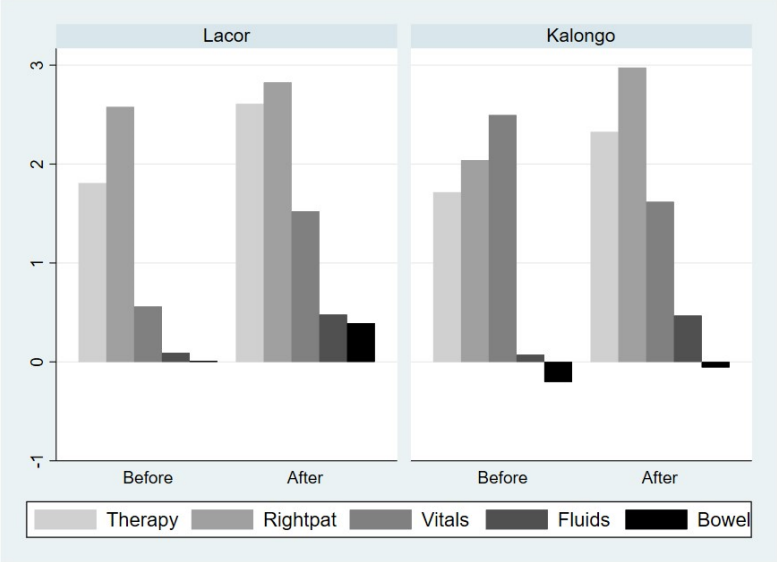
Figure 2 shows an improvement as far as the average scores concerning therapies are concerned. However, the adequacy of blood transfusions (*Blood*) and the correct reporting about the immunization status (*Vacc*) remained in the range [0,1], while the scores attributed to the appropriateness of the treatment (*Treatm*) and the correct use of antibiotics (*Antibio*) showed an upward trend.

Figure 2. Pre-post comparison of therapy items' scores



Notes: *The figure on the left (right) displays the pre-post comparison of the average score attached to each quality item pertaining to therapy procedures at Lacor (Kalongo) hospital.*

Figure 3. Pre-post comparison of nursing items' scores



Notes: *The figure on the left (right) displays the pre-post comparison of the average score attached to each quality item pertaining to nursing procedures at Lacor (Kalongo) hospital.*

Less marked, but still positive, is the variation of the average total score related to nursing procedures (Fig. 3). Overall, the average scores increased with the implementation of the program.

We formally test the hypothesis that the data on quality scores – before and after the implementation of the RBF program – come from different distributions, by performing a battery of Wilcoxon rank-sum tests (e.g., Sprent and Smeeton, 2016).

Table 2: Wilcoxon Rank sum (Mann-Whitney) test

Item	**Before**			**After**			***z***	***Prob* > |*z*|**
	Obs	**ranksum**	**exp.**	**Obs**	**ranksum**	**exp.**		
Symptom	378	86975	113400	221	92724	66300	-14.24	0.000
Exam	378	87112	113400	221	92587	66300	-14.43	0.000
Malaria	378	101054	113400	221	78646	66300	-7.28	0.000
Weight	378	107308	113400	221	72391	66300	-4.26	0.000
Anemia	378	100810	113400	221	78890	66300	-7.55	0.000
Sepsis	378	104979	113400	221	74720	66300	-5.24	0.000
Treatm	378	102060	113400	221	77640	66300	-7.08	0.000
Antibio	378	101477	113400	221	78222	66300	-6.84	0.000
Blood	378	112853	113400	221	66847	66300	-0.33	0.742
Vacc	378	95902	113400	221	83797	66300	-12.84	0.000
therapy	341	83881	93434	208	65997	56444	-6.07	0.000
rightpat	341	84424	93434	208	65454	56444	-7.47	0.000
vitals	341	89376	93434	208	60501	56444	-2.47	0.013
fluids	341	87635	93434	208	62242	56444	-4.42	0.000
bowel	341	91019	93434	208	58858	56444	-2.27	0.023

Notes: *The Wilcoxon rank-sum test checks the hypothesis that two independent samples are from populations with the same distribution. A Prob>\|z\| smaller than the reference value (0.05) indicates that we reject the null hypothesis of no significant difference between the two groups.*

Results, reported in Table 2, show that – except for *Blood* – the null hypothesis that the pre-post samples are drawn from the same population is strongly rejected for all the quality items (p-values are reported in the last column of Table 2).

More specifically, we can say with a high level of confidence that one of the two populations has a significant shift relative to the other.

3.2 *A step forward*

In this Section we deepen the analysis about the effects of the RBF program.

Ideally, we would like to have a treatment and a properly selected statistically equivalent comparison group. With properly designed treatment and control groups, we could attribute all the observable differences in outcomes between treatment and control, to the RBF scheme. As we do not have such Lab-type evidence – at the time the RBF was implemented, it was deemed unfeasible to arrange a larger scale experiment involving a comparison group – we adopt an alternative strategy.

We consider a *spurious treatment group*, which comprises medical and nursing procedures associated with hospitalizations that occurred in a time interval sufficiently close to the inspection days. The *spurious control group* consists of procedures from hospitalizations that took place far from the inspection days (with a distance of more than 14 days before and after any inspection).

What we aim to test is whether the quality of medical procedures in the spurious treatment group is statistically different from that in the spurious control group. This analysis is justifiable because, especially in the context at hand, the medical and nursing procedures implemented during the initial stages of hospitalization play a crucial role in determining the patient's clinical course (Greco et al., 2021).

This approach allows for a more robust examination of the effects of the RBF program. Comparing the spurious treatment and spurious control in 2020 is to be considered more reliable than a simple

pre-post comparison, as different factors may have been at play during the 2014-2016 period but not in 2020 and vice versa. However, it's important to note that this strategy can only detect an effect if the consequences of the treatment are more pronounced in proximity to the inspection days, a point we will address shortly.

In this analysis, we focus on a subset of clinical records from the final year of the project (2020). For each inspection day, we examine two distinct neighbourhoods, each representing a time interval centered around the inspection day, with widths of either fourteen or twenty-eight days. In various regression models, we consider:

a) All the charts compiled either 7 or 14 days before the inspection days.

b) All the charts compiled either 7 or 14 days after the inspection days.

c) All the charts included in neighbourhoods around the inspection days, with dimensions of either fourteen or twenty-eight days.

Each subset of clinical charts falling under a), b), or c) is then compared with all the charts compiled outside a larger period (28 days) centered on the inspection day.

To facilitate this comparison, we utilize a set of dummy variables indicating whether the clinical record falls within the specified neighbourhood of interest or not (please see Table 3 for the definition of these dummy variables).

For each clinical chart, we constructed four quality indicators based on: The sum of all scores related to clinical management, referred to as *Totalscore*; the sum of scores assigned to items evaluating the appropriateness of diagnosis procedures, known as *Diagnosis*; the sum of scores associated with therapy adequacy, labelled as *Therapy*; the sum of all items assessing nurses' performance, denoted as *Nursescore* (refer to Table A1 in the Appendix for detailed information on each item's category).

Table 3: Definition of the dummy variables identifying the clinical charts in the neighbourhood of interest

x_i	x_i=1	x_i = 0
Bef7	Admissions within 7 days before the inspection day	All the charts compiled at a time which lies outside the boundaries of the wider neighbourhood of the inspection day (28 days).
Bef14	Admissions within 14 days before the inspection day	All the charts compiled at a time which lies outside the boundaries of the wider neighbourhood of the inspection day (28 days).
Aft7	Admissions within 7 days after the inspection day	All the charts compiled at a time which lies outside the boundaries of the wider neighbourhood of the inspection day (28 days), excluding those related to patients discharged after the inspection day.
Aft14	Admissions within 14 days after the inspection day	All the charts compiled at a time which lies outside the boundaries of the wider neighbourhood of the inspection day (28 days), excluding those related to patients discharged after the inspection day.
Bef-Aft7	Admissions within 7 days before and 7 days after the inspection day	All the charts compiled at a time which lies outside the boundaries of the wider neighbourhood of the inspection day (28 days).
Bef-Aft14	Admissions within 14 days before and 14 days after the inspection day	All the charts compiled at a time which lies outside the boundaries of the wider neighbourhood of the inspection day (28 days).

Subsequently, we utilized each of these indicators, one at a time, to classify clinical records into three merit classes: *Low, Medium,* and *High*. Specifically, concerning a given indicator, records were

assigned to classes by ranking them according to the scores obtained and then dividing the sample into three equal-frequency groups (see Table 4 for descriptive statistics of the four quality indicators).

Table 4: Descriptive statistics of the quality indicators

Variables	(1) mean	(2) sd	(3) median	(4) min	(5) max	(6) N
Totalscore	24.00	4.41	4.41	11.00	33.00	221
Therapy	8.43	2.77	2.77	1.00	12.00	221
Diagnosis	15.56	2.36	2.36	6.00	21.00	221
Nursescore	10.19	1.35	1.35	7.00	13.00	208

Notes: *The quality indicators are obtained by summing – for any clinical records – all the scores associated to the items pertaining to each category (e.g., diagnosis, therapy, and so on...).*

We conjecture that the clinical charts from different neighbourhoods, particularly those compiled in proximity to the inspection days, will exhibit higher scores compared to those from areas outside the neighbourhoods. This expectation is based on two key reasons.

Firstly, the clinical charts from the neighbourhoods which cover time periods leading up to the inspection days, may reflect increased efforts by health workers who were aware of the impending quarterly evaluation. Given that inspections generally occurred within a two-week window following the close of each quarter, the awareness among health workers of being monitored during those weeks might have motivated them to adhere more rigorously to the best practices under evaluation.

Secondly, the clinical charts from neighbourhoods encompassing time periods following the inspection days could indicate a temporary, heightened commitment among health workers to the feedback received from the committee of experts. In RBF schemes, the regular presence of external supervisors indeed provides an opportunity to assess the actual medical and nursing procedures, which is expected to have a positive impact on these practices.

In performing regression analysis, because the dependent variables of interest are inherently ordered, we use a model widely used for analysing ranking responses, that is, the Ordered Probit model. The

model is built around a latent regression which represents the unobserved structural model underlying the eventual ordinal outcome, that is, the '*intensity of feelings*' leading to the observed ranking.

The latent regression model underlying our ordinal outcomes is identified by the following equation

$$y_i^* = X_i\beta + \epsilon_i \qquad [1]$$

where, for any clinical chart *i*, X_i is the vector containing the full set of explanatory variables, β is the vector of coefficients associated with X_i , and ϵ_i is the disturbance term.

Depending on the specification, y_i denotes one of the quality indicators discussed above (*Totalscore*, *Diagnosis*, *Therapy*, *Nursescore*). The latent regression allows estimating the cutoff parameters μ_h ($h = 1, \dots, 3$), splitting y_i^* into the three observed categories $J \in \{Low, Medium, High\}$.

As far as X_i is concerned, it includes both the dummy variable indicating whether the hospital admission occurred within a specific time interval associated with one of the neighbourhoods centred around the inspection day (see Table 3), as well as a set of additional variables designed to control for other characteristics that could potentially influence quality scores.

Specifically, we consider:

i. children age (in months), *AgeMon*;

ii. the length of hospital stay, *HospStay*;

iii. a dummy variable equal to one whenever the patient was hospitalized at the Lacor hospital (rather than in Kalongo), *Lacor*;

iv. three out of four binary variables identifying the reference quarter, *Quart2*, *Quart3*, *Quart4*, leaving as reference category the first quarter of the year (Jan-Mar). The reference quarter allows to control for differences in the scores that depends on the quarter of implementation of the RBF, as well as seasonality effects.

We then estimate the probability of any chart *i* of falling in each category *J*, according to Equation 2:

$$\Pr(y_i = J|X_i) = \mathrm{F}\,(\mu_h - \mathrm{X_i}\beta) - \mathrm{F}\,(\mu_{h-1} - \mathrm{X_i}\beta) \qquad [2]$$

where F is the normal cumulative density function (cdf), i.e., the probability of the event occurring.

4. Results

Table 5a presents the parameter estimates of the ordered probit regression model (equation 2), with *Totalscore* as the dependent variable. The coefficients for all the dummy variables indicating whether hospital admissions took place in one of the six neighbourhoods around the inspection days are consistently positive and statistically significant. This implies that children who were hospitalized near the inspection days received better care, as evidenced by higher scores for the adequacy of clinical management in their medical records.

Among the control variables, none exhibit statistical significance across all regression specifications. Only the coefficient associated with *AgeMon* remains relatively stable across different model specifications and is statistically significant in most of them. This suggests that as a child's age increases, the quality of clinical management tends to improve.

Our findings do not reveal any discernible differences between the two hospitals, and no significant effect is associated with the quarter in which hospitalization occurred.

To facilitate the understanding of the results, in Table 5*b* we report the predicted probabilities for the three categories of *Totalscore*, and the marginal effects of the dummies identifying the neighbourhoods of the inspection days. The predicted probabilities indicate the Average Adjusted Predictions (AAPs) for each category of score, for hospital admissions occurring within and outside the neighbourhood of interest.

Table 5*a*: Ordered Probit model on *Totalscore* classes (2020)

	(1)	(2)	(3)	(4)	(5)	(6)
Bef7	1.570***					
	(5.66)					
Bef14		1.498***				
		(5.85)				
Aft7			1.421***			
			(4.70)			
Aft14				1.199***		
				(4.93)		
Bef-Aft7					1.503***	
					(7.00)	
Bef-Aft7						1.349***
						(7.29)
AgeMon	0.006	0.007	0.012**	0.009**	0.008*	0.007*
	(1.54)	(1.62)	(2.79)	(2.51)	(2.00)	(1.97)
HospStay	0.062*	0.059*	0.036	0.058	0.05	0.064*
	(2.30)	(2.32)	(0.89)	(1.69)	(1.82)	(2.63)
Lacor	0.228	0.184	0.592*	0.564*	0.159	0.130
	(1.00)	(0.81)	(2.40)	(2.44)	(0.78)	(0.67)
Quart2	-0.172	-0.299	0.142	-0.482	-0.225	-0.854**
	(-0.55)	(-0.96)	(0.38)	(-1.31)	(-0.74)	(-2.67)
Quart4	0.171	0.134	-0.797	-0.673	-1.050	-1.012
	(0.55)	(0.44)	(-1.28)	(-1.22)	(-1.69)	(-1.77)
cut1	0.633*	0.571*	0.936***	0.889***	0.569*	0.509*
	(2.38)	(2.18)	(3.39)	(3.36)	(2.30)	(2.20)
cut2	1.796***	1.754***	2.082***	1.983***	1.692***	1.613***
	(6.46)	(6.47)	(6.83)	(6.89)	(6.27)	(6.49)
N†	135	140	131	143	167	184

Notes: *t statistics in parentheses. Statistical significance * p < 0.05, ** p < 0.01, *** p < 0.001; Quart3 omitted because of lack of observations. †The number of observations varies across the different specifications due to the observations excluded by the definition of the dummies identifying the neighbourhoods surrounding the inspection.*

Table 5b: Predicted probabilities and marginal effects from the estimated ordered probit on *Totalscore* classes

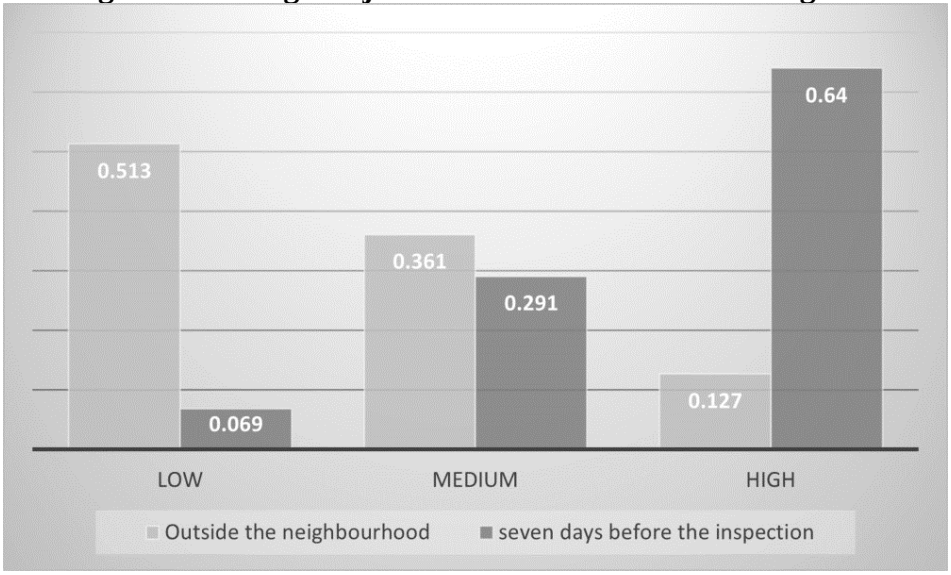
A) Predicted probabilities		**Low**	**Medium**	**High**
Bef7	0	0.513	0.361	0.127
	1	0.069	0.291	0.64
Bef14	0	0.502	0.369	0.129
	1	0.076	0.308	0.617
Aft7	0	0.529	0.343	0.128
	1	0.106	0.322	0.572
Aft14	0	0.529	0.331	0.14
	1	0.154	0.337	0.51
Bef-Aft7	0	0.524	0.348	0.128
	1	0.083	0.297	0.619
Bef-Aft14	0	0.512	0.343	0.145
	1	0.112	0.311	0.578

B) Marginal effects	**Low**	**Medium**	**High**
Bef7	-0.444***	-0.07	0.513***
Bef14	-0.426***	-0.061	0.487***
Aft7	-0.423***	-0.021	0.444***
Aft14	-0.375***	0.006	0.369***
Bef-Aft7	-0.441***	-0.05	0.491***
Bef-Aft14	-0.401***	-0.0321	0.433***

Notes: *Statistical significance :* * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Results suggest that, on an *all other things equal* basis, the clinical records concerning hospital admissions which occurred within 7 days before the inspections, are more than five times as likely as those outside the largest neighbourhood to be in the *High* score class (64% as opposed to 12.7%, see Figure 4).

Figure 4. Average Adjusted Predictions for score categories



Notes: *The figure shows the Average Adjusted Predictions (AAPs) for each category of score, for hospital admissions occurring within 7 days before the inspection day and those outside that neighbourhood.*

Similar findings are observed when examining various time intervals within the neighbourhoods centred on the inspection day, such as Bef14, Aft7 (or alternatively, 14), and Bef-Aft7 (or alternatively, 14), as shown in column 3 of Table 5b. Conversely, if we analyse the predicted probability of being in the *Low* score category, this probability is higher for children who were not hospitalized during any of the intervals within the relevant time neighbourhoods.

Section B of Table 5b presents the marginal effects, which indicate the change in the probability of belonging to one of the three categories for discrete changes in the dummy variables indicating hospitalization in proximity to the inspections. The results suggest an increase in the probability of falling into the *High* score category, which ranges from 43% to 51%, as a result of being hospitalized

near the inspection days. These results are symmetrical concerning the *Low* score category, while no statistically significant differences emerge in the probability of belonging to the *Medium* score category based on the timing of hospitalization (whether near the inspection days or not). We also consider wider neighbourhoods to check whether the difference between clinical records in the *spurious treatment group* and those in the *spurious control group* disappears. More precisely, we enlarge the neighbourhood to both 21 and 28 days. Results, available upon request, confirm the difference between *treatment* and *control*. Remarkably, they show a reduction in the size of the coefficients. As an example, the increase in the predicted probability of being in the *High* score category fall from 64% (57%) to 52% (48%) when we move from clinical records managed within 7 days before (after) the inspection to those managed before (after) 28 days. This suggests that the effect soften as we move away from the inspection days.

Table A2 in the Appendix provides a more detailed examination of the factors contributing to the observed increase in *Totalscore*. When it comes to *Diagnosis* (Columns 1-3) and *Therapy*, it presents the marginal effects of belonging to any score class. Interestingly, in the case of *Diagnosis*, changes in the probability of being assigned to a specific score class are not statistically significant for almost all of the considered time intervals. An exception to this is the broadest neighbourhood, which compares clinical charts from the 28 days surrounding the inspection days with those from other time periods. In this case, we observe an increase in the probability of belonging to the *High* score class (20.8%), and a symmetrical decrease in the probability of belonging to the *Low* score class (-17.4%). Regarding the appropriate administration of therapies (*Therapy* score), the marginal effects are statistically significant and display the expected pattern. Specifically, there is an increase, ranging from 47.1% to 75.9%, in the probability of belonging to the *High* score class for clinical charts related to children hospitalized within 7 (or alternatively, 14) days before and after the inspection.

This evidence indicates that the overall score increase is primarily driven by enhancements in the appropriateness of treatments provided.

Table SM1 in the Supplementary Material presents the results related to the estimation of the ordered probit model in equation 2, where the dependent variable is *Nursescore* (the sum of all items assessing nurses' performance). The coefficients associated with the dummy variables that capture the impact of being hospitalized in a specific time interval within the neighbourhoods around the inspection days are consistently found to be statistically insignificant.

To test the robustness of our findings, we conducted two additional tests. First, we estimated a simple OLS model. The rationale behind this test is to consider whether the results hold when directly considering scores instead of classes. The results, which remain consistent with our earlier findings, are reported in Table SM2 in the Supplementary Material.

Second, we conducted a placebo population test. This test is informative for assessing bias, particularly if there are reasons to believe that the assumed treatment effect does not apply to the placebo population, while any potential flaws would operate similarly (e.g., Eggers et al., 2021). In this test, we examined a sample of 367 clinical charts related to children hospitalized before the commencement of the RBF program (years 2014-2016). We replicated the same ordered probit model from equation 2, using score classes derived from *Totalscore* as the dependent variable. In estimating the model, we assumed that the inspections occurred on the same days as in 2020. This assumption helps eliminate the possibility that our primary results were influenced by factors not accounted for in the regression model, such as seasonality. The results are presented in Table SM3 in the Supplementary Material. There were no statistically significant differences in *Totalscore* between the clinical charts from any of the relevant neighbourhoods around the inspection days and those from other time periods.

It is noteworthy that differences between hospitals are now statistically significant. While there is evidence that the RBF program had a greater impact on the hospital that was initially lagging behind (Greco et al., 2021), it cannot be definitively concluded that these differences were blurred by the RBF program. It's plausible that Kalongo Hospital made independent advancements in the quality of care that are not directly attributable to the program.

Our results indicate that nurses appeared to be less responsive to the evaluations conducted on inspection days. Therefore, the overall improvement observed in the scores assessing their performance during the RBF program seems to have followed a more consistent pattern throughout the entire year.

5. Discussion and conclusions

In past decades, the Results-Based Financing approach has spread out in many developing countries as a novel organizing model to finance the provision of health care. As a potential response to the growing budget pressure and the increasing attention to achieving measurable results, this method has been implemented with the aim of improving both healthcare quality and the utilization of health services.

However, despite an increasing body of evidence suggests that this approach might increase both quantity and quality of healthcare in developing countries if the incentive scheme is carefully designed, some authors have emphasized its potential side-effects.

In a behavioral perspective, the mere awareness of being supervised and audited might indeed play a crucial role in changing the performance of employees, incentivizing staff compliance to good practices (Chen et al., 2015); supervision is supposed to enhance motivation especially if perceived as a means to support performance, rather than simply as a device to control activity (Lemière et al., 2013; Frey and Jegen, 2001).

In this paper we have assessed a RBF project targeted at the children wards of two Ugandan hospitals by relying on data from a randomly extracted sample of clinical records.

Our findings are consistent with a dual effect of the supervision process on the quality scores that assess the clinical management of patients.

First, we observed an increase in the probability of receiving a score assessing the appropriateness of clinical management for the medical records managed within 7 (14) days before the inspection. This aligns with the Hawthorne effect, which suggests a positive impact on health workers' performance due to their awareness of being supervised. Second, our results show an increase in the probability of receiving a score in the highest class, ranging from 37% to 44%, for the clinical records handled within two weeks after the inspection. This is consistent with the literature suggesting a positive impact of supervision when it is perceived as supportive rather than controlling (Lemière et al., 2013; Frey and Jegen, 2001). In this context, the audit serves as an opportunity to share information about the project objectives.

These results pertaining to clinical management are primarily driven by significant improvements in therapy procedures.

Since no evidence is found regarding the appropriateness of nursing care in proximity to the inspection days, following the promising results related to clinical management scores, it may be advantageous to design RBF schemes that facilitate a broader involvement of nurses in achieving the project objectives.

The present study comes with several limitations that should be addressed in dedicated analyses. Firstly, it is challenging to definitively determine whether the effects we observe on the quality of clinical procedures are primarily driven by monetary incentives linked to good performance or if they are simply a consequence of monitoring. There may be psychological or reputational effects at play that are, unfortunately, not detectable in this study.

Secondly, while our study represents progress in the right direction, it is only fair to acknowledge the

need for more robust evidence concerning the persistence of the effects generated by RBF programs. Investigating these potential long-term effects could be particularly valuable in assessing the degree to which the best practices established during the pediatric RBF program have become ingrained in the behavior of healthcare workers.

References

Bitton, A., Fifield, J., Ratcliffe, H., Karlage, A., Wang, H., Veillard, J. H., Schwarz, D., & Hirschhorn, L. R. (2019). Primary healthcare system performance in low-income and middle-income countries: a scoping review of the evidence from 2010 to 2017. BMJ Global Health, 4(Suppl 8), e001551.

Bowles, S. (2008). Policies designed for self-interested citizens may undermine" the moral sentiments": Evidence from economic experiments. *Science*, 320(5883), 1605-1609.

Brandts, J., & Cooper, D. J. (2007). It's what you say, not what you pay: An experimental study of manager-employee relationships in overcoming coordination failure. *Journal of the European Economic Association*, 5(6), 1223-1268.

Chen, L. F., Vander Weg, M. W., Hofmann, D. A., & Reisinger, H. S. (2015). The Hawthorne effect in infection prevention and epidemiology. *Infection control & hospital epidemiology*, 36(12), 1444-1450.

Di Pietro, L., Piaggio, D., Oronti, I., Maccaro, A., Houessouvo, R. C., Medenou, D., De Maria, C., Pecchia, L., & Ahluwalia, A. (2020). A framework for assessing healthcare facilities in low-resource settings: field studies in Benin and Uganda. Journal of Medical and Biological Engineering, 40, 526-534.

Eggers, A. C., Tuñón, G., & Dafoe, A. (2021). Placebo tests for causal inference. Unpublished manuscript. https://pelg. ucsd. edu/Eggers_2021. pdf.

Eldridge, M., & TeKolste, R. (2016). Results-based financing approaches. Washington DC: Urban Institute.

Fretheim, A., Witter, S., Lindahl, A. K., & Olsen, I. T. (2012). Performance-based financing in low- and middle-income countries: still more questions than answers. Bulletin of the World Health Organization, 90, 559-559A.

Frey, B. S., & Jegen, R. (2001). Motivation crowding theory. Journal of economic surveys, 15(5), 589-611.

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. Journal of economic perspectives, 25(4), 191-210.

Greco, L., Ochola, E., Omona, V., Okot, G.S., & Mozzi, V (2021). Results based financing, an engine of change for pediatric services, 2021, from https://fondazionecorti.it/wp-content/uploads/2014/07/RBF-study-2021.pdf

Grittner, A. M. (2013). Results-based Financing: Evidence from performance-based financing in the health sector. Discussion Paper No. 6/2013. German Institute of Development and Sustainability (IDOS), Bonn.

Harrington, J. (2018). From the Solomon Islands to Liberia: These are the 25 poorest countries in the world. USA Today. Retrieved May 5, 2020, from https://eu.usatoday.com/story/money/2018/11/29/poorest-countries-world-2018/38429473/.

Honda, A. (2013). 10 best resources on… pay for performance in low-and middle-income countries. Health policy and planning, 28(5), 454-457.

Laffont, J. J., & Martimort, D. (2009). The theory of incentives: the principal-agent model. In The theory of incentives. Princeton university press.

Lemière, C., Torsvik, G., Mæstad, O., Herbst, C. H., & Leonard, K. L. (2013). Evaluating the Impact of Results-Based Financing on Health Worker Performance: Theory, Tools and Variables to Inform an Impact Evaluation, *Health, Nutrition and Population (HNP) Discussion Paper Series* 98269, The World Bank.

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. Journal of mathematical sociology, 4(1), 103-120.

Mejía-Guevara, I., Zuo, W., Bendavid, E., Li, N., & Tuljapurkar, S. (2019). Age distribution, trends, and forecasts of under-5 mortality in 31 sub-Saharan African countries: A modeling study. PLoS medicine, 16(3), e1002757.

Mushasha, R., & El Bcheraoui, C. (2023). Comparative effectiveness of financing models in development assistance for health and the role of results-based funding approaches: a scoping review. Globalization and Health, 19(1), 1-11.

Oxman, A. D., & Fretheim, A. (2008). An overview of research on the effects of results-based financing. Norwegian Knowledge Centre for the Health Services.

Rowe, A. K., De Savigny, D., Lanata, C. F., & Victora, C. G. (2005). How can we achieve and maintain high-quality performance of health workers in low-resource settings?. The Lancet, 366(9490), 1026-1035.

Savedoff, W. D., & Partner, S. (2010). Basic economics of results-based financing in health. Bath, Maine: Social Insight.

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, 1-15.

Shroff, Z. C., Bigdeli, M., & Meessen, B. (2017). From scheme to system (part 2): findings from ten countries on the policy evolution of results-based financing in health systems. Health Systems & Reform, 3(2), 137-147.

Sprent, P., & Smeeton, N. C. (2016). Applied nonparametric statistical methods. CRC press.

Ssengooba, F., Ekirapa, E., Musila, T., & Ssennyonjo, A. (2015). Learning from multiple results-based financing schemes: an anlysis of the policy process for scale-up in Uganda (2003–2015). Geneva, Switzerland: Alliance for Health Policy and Systems Research, WHO.

Titmuss, R. M. (1970). *The gift relationship*. London: Allen & Unwin.

United Nations, (2011). *Results-based Management Handbook*. Available on-line at https://unsdg.un.org/sites/default/files/UNDG-RBM-Handbook-2012.pdf

Zikusooka, C. M., Kyomuhang, R., Orem, J. N., & Tumwine, M. (2009). Is Health Care Financing in Uganda Equitable? *African Health Sciences*, 9(2).

Appendix

Table A1: Checklist of the quality items

Items	Description	Obs	
		2014-16	2020
*A) **Clinical management***			
Diagnosis			
Symptom	Clinical history	378	221
Exam	Clinical examination	378	221
Malaria	Malaria excluded or treated (fever)	378	221
Weight	Measuring of weight	378	221
Perc	Percentile charts available	378	221
Anemia	Anemia diagnosed	378	221
Sepsis	Sepsis specific diagnosis	378	221
Diagnosis	Sum of Symptom, Exam, Malaria, Weight, Perc, Anemia Sepsis	378	221
Therapy			
Treatm	Appropriate treatment	378	221
Antibio	Antibiotics only if necessary	378	221
Blood	Appropriate request of blood transfusions	378	221
Vacc	Check vaccination record	378	221
Therapy	Sum of Treatm, Antibio, Blood, Vacc	378	221
Clinical	Clinical management: sum of Symptom, Treatm, Exam, Weight, Antibio	378	221
Totalscore	Sum of all clinical management items in A)	378	221
*B) **Nursing procedure***			
therapy	Proper therapy administration	341	208
rightpat	Charts conformity to patients	341	208
vitals	Reporting of weight and vital signs	341	208
fluids	Presence of fluid balance chart	341	208
bowel	Recording of bowel events	341	208
Nursescore	Nursing procedures' total score: sum of all items in B)	341	208

Notes: *The table reports the comprehensive list of quality items used to assess the quality of both clinical (Panel A) and nursing procedures (Panel B).*

Table A2: Marginal effects from the estimated ordered probit on quality score classes

	Diagnosis score			*Therapy score*			
	(1)	(2)	(3)	(4)	(5)	(6)	
Marginal effect	**Low**	**Medium**	**High**	**Low**	**Medium**	**High**	Obs
	6-14	15	16-21	1-6	7-9	10-12	
Bef7	-0.126	-0.017	0.144	-0.277***	-0.482***	0.759***	135
Bef14	-0.141	-0.021	0.162	-0.272***	-0.430***	0.701***	140
Aft7	0.124	-0.023	0.148	-0.286***	-0.403***	0.689***	131
Aft14	-0.124	-0.034	0.196*	-0.272***	-0.199**	0.471***	143
Bef-Aft7	-0.159*	-0.029	0.188*	-0.285***	-0.398***	0.683***	167
Bef-Aft14	-0.174**	-0.034*	0.208**	-0.277***	-0.266***	0.543***	184

***Notes**: Statistical significance: *** p<0.01, ** p<0.05, * p<0.1. The figures below the classes show the min-max values of score for each class.*

Supplementary material

Table SM1: Ordered Probit model on *Nursescore* classes (2020)

	(1)	(2)	(3)	(4)	(5)	(6)
Bef7	0.091					
	(0.28)					
Bef14		0.171				
		(0.54)				
Aft7			0.810			
			(1.26)			
Aft14				-0.245		
				(-0.95)		
Bef-Aft7					0.401	
					(1.45)	
Bef-Aft7						0.079
						(0.39)
AgeMon	0.003	0.003	0.003	0.004	0.003	0.003
	(0.80)	(0.87)	(0.93)	(1.38)	(0.96)	(1.27)
HospStay	0.05*	0.052*	0.037	0.04	0.043	0.049*
	(2.01)	(2.12)	(1.46)	(1.76)	(1.79)	(2.22)
Lacor	0.130	0.164	0.305	0.311	0.177	0.0530
	(0.47)	(0.59)	(1.04)	(1.07)	(0.65)	(0.20)
Quart2	-0.139	-0.086	0.087	0.038	-0.119	-0.337
	(-0.47)	(-0.30)	(0.27)	(0.12)	(-0.41)	(-1.35)
Quart3	0.120	0.130	0.467	0.327	0.0374	-0.263
	(0.31)	(0.33)	(1.09)	(0.89)	(0.10)	(-0.92)
Quart4	0.009	0.02	0.286	0.434	0.313	0.237
	(0.02)	(0.05)	(0.70)	(1.11)	(0.88)	(0.68)
cut1	-0.228	-0.175	-0.142	-0.055	-0.182	-0.242
	(-0.78)	(-0.61)	(-0.47)	(-0.19)	(-0.65)	(-0.97)
cut2	0.845**	0.893**	1.023**	1.017***	0.866**	0.746**
	(2.84)	(3.04)	(3.26)	(3.42)	(3.02)	(2.94)
N	136	137	123	157	146	181

***Notes**: t statistics in parentheses. Statistical significance: * p < 0.05, ** p < 0.01, *** p < 0.001*

Table SM2: Linear relationship between audit and *Totalscore*

	(1)	(2)	(3)	(4)	(5)	(6)
Constant	19.67***	19.93***	19.05***	19.09***	19.91***	20.16***
	(0.943)	(0.906)	(0.962)	(0.902)	(0.814)	(0.749)
Bef7	5.359***					
	(0.77)					
Bef14		5.025***				
		(0.727)				
Aft7			4.934***			
			(0.857)			
Aft14				4.620***		
				(0.724)		
Bef-Aft7					5.234***	
					(0.609)	
Bef-Aft14						4.857***
						(0.545)
AgeMon	0.031*	0.029*	0.042***	0.033**	0.034**	0.026**
	(0.013)	(0.011)	(0.012)	(0.011)	(0.012)	(0.01)
HospStay	0.155	0.135	0.118	0.206	0.127	0.156*
	(0.08)	(0.0741)	(0.135)	(0.111)	(0.075)	(0.063)
Lacor	0.862	0.719	1.508	1.561	0.641	0.675
	(0.817)	(0.795)	(0.854)	(0.807)	(0.666)	(0.626)
Quart2	0.00548	-0.255	0.696	-0.979	-0.154	-1.735*
	(0.873)	(0.842)	(0.974)	(0.959)	(0.820)	(0.806)
Quart4	-0.528	-0.469	-2.873*	-2.504*	-3.440*	-3.104*
	(1.011)	(0.945)	(1.364)	(1.195)	(1.323)	(1.207)
Observations	135	140	131	143	167	184
$x_i = 1$	26	31	22	34	58	75
R^2	0.298	0.308	0.247	0.279	0.351	0.358
Adjusted R^2	0.266	0.277	0.210	0.247	0.326	0.336

Notes: *Robust standard errors in parentheses. Statistical significance: *** p<0.01, ** p<0.05, * p<0.1. Quart3 omitted because of lack of observations.*

Table SM3: Placebo Test - Ordered Probit model on *Totalscore* classes (2014-2016)

	(1)	(2)	(3)	(4)	(5)	(6)
Bef7	0.192					
	(0.58)					
Bef14		0.192				
		(0.58)				
Aft7			-0.325			
			(-1.51)			
Aft14				-0.342		
				(-1.70)		
Bef-Aft7					-0.154	
					(-0.87)	
Bef-Aft7						-0.182
						(-1.07)
AgeMon	0.003	0.003	0.004	0.004*	0.004	0.004*
	(1.69)	(1.69)	(1.83)	(1.96)	(1.82)	(1.97)
HospStay	-0.002	-0.002	-0.002	-0.002	-0.0021	-0.002
	(-0.68)	(-0.68)	(-0.75)	(-0.75)	(-0.71)	(-0.71)
Lacor	0.361**	0.361**	0.365**	0.368**	0.347**	0.354**
	(2.69)	(2.69)	(2.73)	(2.79)	(2.69)	(2.77)
Quart2	-0.110	-0.110	-0.146	-0.161	-0.150	-0.163
	(-0.16)	(-0.16)	(-0.22)	(-0.24)	(-0.22)	(-0.24)
Quart3	0.263	0.263	0.007	-0.014	-0.011	-0.027
	(1.16)	(1.16)	(0.03)	(-0.07)	(-0.05)	(-0.14)
cut1	-0.304*	-0.304*	-0.319**	-0.310**	-0.333**	-0.320**
	(-2.51)	(-2.51)	(-2.64)	(-2.59)	(-2.82)	(-2.74)
cut2	0.625***	0.625***	0.609***	0.615***	0.599***	0.607***
	(5.06)	(5.06)	(4.96)	(5.05)	(4.99)	(5.09)
N	326	326	347	352	362	367

Notes: *t statistics in parentheses. Statistical significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.*