REVISED WHP EVALUATION DESIGN NOTE (June 2013)

(1) Questions that can still be addressed with rigor

The originally planned experimental design for the World Health Partners (WHP) evaluation is no longer feasible (implementation levels in "control" areas and "treatment" areas are almost identical). However, the high quality household baseline data from a large representative sample of households from 360 study clusters combined with data from the upcoming follow-up round enables us to address all of the central study questions originally proposed (although the research questions related to TB and VL have been removed from the study). We therefore focus our attention largely on developing an appropriate non-experimental design and describing its strengths and limitations, but we first briefly review the questions that can be addressed.

RESEARCH QUESTIONS:
(A) Primary outcomes:
Impact of the WHP program on the target disease outcomes (originally requested by the Gates Foundation): These outcomes focus on receipt and completion of therapies known to be efficacious, as well as delays in treatment initiation and duration of illness, for major childhood illnesses:

- Diarrhea: Proportion of children under 5 with diarrhea in last 2 weeks who were treated with zinc
- Pneumonia: Proportion of children under 5 with suspected pneumonia in last 2 weeks who received a full course of antibiotics (five days)
- Number of days that a sick child remains symptomatic (duration of illness) (for both diarrhea and pneumonia)
- Time between onset of symptoms and seeking care for a sick child (for both diarrhea and pneumonia)

(B) Quality of care:
Impact of WHP's programs on provider knowledge and quality of medical care received for childhood diarrhea and pneumonia: These are key pathways through which primary outcomes might be improved.

- Provider Knowledge: Assessed using Vignettes
- Quality of Medical Care: Assessed using Standardized Patients

(C) Health service utilization, health spending, and WHP cost structure:
Impact of improvements in quality on demand (healthcare utilization) and on households' healthcare expenditures: These are factors related to sustainability and longer-term impact including the overall cost of WHP program at which the improvements are achieved.

- Number of days of work missed by caretakers/forgone income
- Total out-of-pocket expenditures on medical care, total expenditure on home-remedies or self-prescribed medicines, and overall budget share devoted to illness-related expenses

- Total and marginal costs of service delivery by WHP and how these relate to total service volume and time/experience in providing services

(2) The intended design (including alternatives) now that the experimental design is no longer feasible.

The theory of change underlying the WHP model is that improvements in service quality (through technology, training, standardization, and supply-chain engineering) coupled with demand-generation (to reach a substantial share of the Bihar population with higher quality services) will improve primary outcomes. As with the original experimental design, provider and household surveys enable us to distinguish supply-side improvements (service quality), demand-side changes (service use/market share), and how they combine to influence study outcomes.

However, we propose a revised study design that takes into account the fact that study areas where WHP implementation has happened might be different from those in which it did not. Given the richness of our baseline data, we propose to use a methodology known in economics as difference-in-differences (DD) and in epidemiology as controlled before-after design.[1] In brief, this approach isolates differential changes over time (between baseline and follow-up) in outcomes of interest between implementation and non-implementation study clusters, attributing changes in outcomes to implementation of the program. *(In attached appendices, we provide several of our papers as examples of high-impact applications of this methodology, and we also provide an extended analysis of the assumptions required for this approach to be sound showing that these assumptions hold for the WHP evaluation).*

Under plausible assumptions (which we explain and assess below), this DD method takes into account pre-existing differences between implementation and non-implementation clusters by comparing *changes* in the outcome variables rather than the *levels* of the outcome variables (as would be possible in a randomized controlled trial). For example, if the *increase* in the proportion of children with diarrhea who are treated with zinc is 0.2 in implementation areas but only 0.05 in non-implementation areas (after accounting both for a observed potential confounders as well as unobserved confounders that remain fixed over time), then WHP program will be estimated to have an impact of 0.15 on this outcome.

More formally, to estimate the impact of WHP's programs on an individual level outcome *y* (for example, whether or not a child with diarrhea was treated with zinc), we estimate the following basic regression equation:

$$y_{ihct} = \alpha_0 + \alpha_1 WHP_{ct} + \alpha_2 X_{ihct} + \alpha_3 K_{ct} + \gamma_c + \theta_t + \varepsilon_{ihct}, \quad t = 1,2$$

where y_{ihct} refer to the value of the outcome variable of interest for individual *i* living in household *h* in cluster *c* at time *t* (=1 for baseline and 2 for follow-up), WHP_{ct} takes

[1] Cochrane Review: http://ccg.cochrane.org/non-randomised-controlled-study-nrs-designs

value 1 in the WHP program has been implemented in cluster *c* in time *t* and 0 if not, X_{ihct} are individual *i* and household *h* observable characteristics which are recorded in the rich dataset that has and will be collected (dwelling characteristics, assets, socio-economic variables), K_{ct} refers to any cluster-level time-varying characteristics that are measurable (e.g., differences in the total population/total number of households in villages enumerated at baseline and study follow-up), γ_c are cluster-level fixed effects that control for time invariant characteristics of each cluster, and θ_t are time period fixed effects that control for unobservable shocks that affect all households surveyed in each period. The parameter of interest, which estimates the impact of the WHP intervention on the outcome variable y, is α_1.

This methodological approach can also easily be adapted to analyze how the WHP program differentially impacts households with certain characteristics. Of particular interest are Below Poverty Line (BPL) households because their fees for WHP consultations are fully subsidized. In consultation with WHP and the Gates Foundation, we will develop a list of other important household characteristics by which we can assess heterogeneous program impact.

Our study design can also be used to estimate changes in provider knowledge (measured using vignettes) and quality of care provided in study clusters (measured using Standardized Patients) under WHP's programs with estimating equations of the following basic form:

$$y_{pct} = \alpha_0 + \alpha_1 WHP_{ct} + \alpha_2 X_p + \alpha_3 K_{ct} + \gamma_c + \theta_t + \varepsilon_{ihct}$$

where y_{pct} refer to provider knowledge or service quality for provider *p* operating in cluster *c* at time *t* (1 for follow-up and 0 for baseline), WHP_{ct} takes value 1 in the WHP program has been implemented in cluster *c* in time *t* and 0 if not, , X_{pct} are provider *p* time invariant characteristics (year of birth, qualifications and basic facility characteristics at the time of baselines) K refer to refers to any cluster-level measurable time-varying characteristics, γ_c are cluser-level fixed effects absorbing time invariant characteristics of each cluster, and θ_t are time period fixed effects to control for unobservable shocks that affect all providers surveyed in each period.

Although improvements in provider knowledge and quality of care are central pathways through which WHP might improve the treatment of childhood diarrhea and pneumonia, it is important to distinguish demand- and supply-side effects in evaluating the WHP program. For example, even if WHP effectively improved the quality of services available through affiliated providers, this will do little to improve appropriate treatment in local populations if local residents do not use WHP-affiliated provider services at high rates relative to other providers (i.e., if WHP does not have a large market share). Our design and use of provider quality assessments together with household surveys allows us to distinguish these demand- and supply-side factors (both on average for the entire study population and for BPL households who pay lower prices when visiting WHP providers and for whom effects could be more likely). In doing so, the evaluation will be able to

identify if further improvements are better focused on improving service quality or on increasing coverage (or market share) in local populations.

We also highlight that there are ways that the WHP program could influence population health without working through provider knowledge and quality. WHP operates a large-scale distribution program for of affordable and quality-assured drugs. Because these medicines are often cheaper than market alternatives, compliance with treatments may improve. WHP has also introduced a network of inexpensive, quality-assured lab tests that may increase diagnostic capabilities in ways that cannot be captured by the Standardized Patient methodology. Our household survey will be able to detect changes that occur through all of these pathways, and they will also allow us to study the anatomy of how households choose providers (with potential insights for how WHP could expand its coverage of local populations/market share).

Relative to the original experimental design, the main limitation of the difference-in-difference approach that we propose is that unobserved time-varying factors may still play a confounding role (the DD method allows for unobserved potential confounders that are fixed over time, and the rich data we have from baseline and follow-up allows us to observe and account for a broad set of time-varying candidate confounders). There are well-accepted practices in applied statistics for assessing the extent to which the assumptions of the DD method does not pose interpretational problems. Specifically, we can test the degree to which the implementation of WHP's programs is correlated with pre-existing trend differences in child health and survival outcomes. Doing so requires data on child health outcomes collected at multiple points in time prior to program implementation. Because our baseline data set is a single cross-section, we use data from the District-level Household and Facility Surveys (DLHS) to test this 'parallel trends' assumption. Importantly, when we examine these 'parallel trends', we find little evidence of systematic correlations between intensity of WHP program implementation and pre-existing trends infant and neonatal mortality. These details are shown Appendix 1.

We also use our baseline evaluation data to test for differences in characteristics between study clusters in which WHP has and has not implemented programs to date. The advantage of doing so relative to the DLHS analysis is that we are able to make these comparisons at the study cluster level. The disadvantage is that we are only able to test for differences in child health *levels* rather than trends between implementation and non-implementation areas. We do not find any evidence of broad level differences between implementation and no implementation clusters in either key study outcomes or related variables, even without applying the appropriate multiple comparison corrections. Like the DLHS analysis of pre-existing district trends, this cluster-level analysis suggests that implementation is not strongly linked to differences in levels of the key study outcomes and that a difference-in-difference design is a reasonable approach for the evaluation.

Finally, we note that we have extensive experience conducting impact evaluations using this methodological approach, both in India and elsewhere. We attach two examples in an appendix (one is a recent evaluation not yet published of the Gujarat's Chiranjeevi

Scheme, the other is an evaluation of President's Emergency Plan for AIDS Relief (PEPFAR) in Africa published in the *Journal of the American Medical Association*).

(3) Revised timeline

The *ideal* time to collect the follow-up data is May-October 2014 to coincide with the timing of the baseline, making the two waves comparable in terms of seasonal patterns of weather and disease. Moreover, this will give enough time for WHP to further their program, as current implementation is still behind its goals. According to our monitoring data, WHP is currently implementing in a third of the study clusters (131 clusters, as of June 19, 2013). Based on our communication with WHP, the MiniRemedi application has suffered some technical challenges but they anticipate that it will be ready for distribution in 3 to 4 months from now. Consequently, by May 2014 the WHP program is likely to be operating in approximately 40%-60% of the clusters.

However, given the urgency that the Foundation and WHP face for evidence on the impact of WHP program, and possible mechanisms, it is possible to conduct the follow-up data collection earlier (as early as December 2013, for example), although this strategy would create challenges with seasonal variation. The DD strategy accommodates such differences across the implementation and non-implementation areas, it is important to note that diarrheal diseases and pneumonia are prone to seasonal variation.

With the baseline study data, cleaning, and coding the data along with addressing issues of matching household data with provider data took several additional months. Based on this experience and planned modifications in the data collection, we expect to have a clean dataset from the follow-up round ready for analysis in 3 months after field work is completed. We expect to present analysis on main research questions in the next 3 months, and draft manuscripts to be circulated for comments in 6 months after receipt of data. This timeline is estimated based on availability of research support at levels requested in our supplement budget.

<u>Assuming data collection starting December 2013</u>:

Preparation work (modification of questionnaires, protocols):	Aug-Dec 2013
Completion of Household Data Collection:	April 2014
Completion of Provider Data Collection:	May 2014
Datasets ready for analysis:	July 2014
Preliminary findings presented at MLE:	October 2014
Draft Manuscripts for comments:	January 2015
Full set of papers presented at MLE:	March 2015
All papers submitted, and dataset transferred to the foundation:	July 2015

<u>Assuming data collection starting May 2014</u>:

Preparation work (modification of questionnaires, protocols):	Jan-May 2014
Completion of Household Data Collection:	September 2014
Completion of Provider Data Collection:	October 2014
Datasets ready for analysis:	December 2014

Preliminary findings presented at MLE:	February 2015
Draft Manuscripts for comments:	May 2015
Full set of papers presented at MLE:	August 2015
All papers submitted, and dataset transferred to the foundation:	December 2015

(4) Components that could be included or not (and their timing/cost)

We envision including all the key components from the baseline (house listings, main household survey, provider surveys, vignettes, and standardized patients) except the household surveys dedicated to TB and VL, and the network component. The Gates Foundation has determined that the TB and VL components will not be carried out. The network component does not seem appropriate after the WHP changed its business model by deemphasizing the importance of the referral process between Rural Health Providers (RHPs) and Telemedicine Providers (TPCs). The analysis reported above using DLHS and the BEST baseline data give us confidence that we can rigorously answer the original questions using the controlled before-after design (i.e., difference-in-difference design), which is why we propose to include most of the original components.

An important option to conserve resources is to use the December 2013 timeline for the follow-up, rather than starting in May 2014. This would shorten the overall length of the project, and help reduce personnel costs.