Reviewer's report

Title: Can we trust measures of healthcare utilisation from household surveys?

Version: 2 Date: 18 June 2013

Reviewer: Jishnu Das

Reviewer's report:

Summary: The paper compares morbidity and health seeking behavior among children in Ghanaian households using two different measures: Pictorial diaries, where care-givers record morbidity when an episode occurs, and household surveys that ask about the child’s most recent illness without a pre-specified recall period. The measures differ. The authors argue that the household survey undercounts morbidity (for everyone) and utilization (more so among the poor) and that depending on the source of the data, the impact of a randomized trial on utilization changes.

Comments: I am very sympathetic to the general program that is advanced in this paper—that we do not sufficiently understand how health care events are recalled and reported in household survey, and that changes in various dimensions of the survey can have large impacts on basic measures of morbidity and health seeking. The writing is clear, and the results are believable. I have some suggestions that the authors might wish to consider for a revised version of their paper.

Main Suggestion:

I. Measures of Morbidity: My main suggestion is clarifying precisely what was done, and what can be understood from the data. The suggestion is motivated, in part, by the difficulty of parsing out the duration of recall and potential seasonal effects from the impact of the data source. Here is what I was thinking:

Imagine that you have two time periods, and no seasonal effects, so that morbidity and health seeking are identical in both. The `true' data is that, in a population of 100 families, 20 have an illness only in T1, 20 have an illness only in T2 and 40 have an illness in both T1 and T2. With no biases, the household survey will then pick up the following report:

20 households will report an illness in T1
60 households will report an illness in T2, since this is their most `recent' illness.

As far as I understand it, the pictorial diary will actually pick up 120 illness episodes (20 + 20 + 40*2). The difference of 40 comes because the 40 families who had an illness in both T1 and T2 will report only the T2 illness in the household survey. Therefore, the impact of the method on total illness episodes cannot be determined. What can be determined is the number of households not reporting an illness episode in T2 and the number of households not reporting an
illness episode in either T1 or T2. These are 40 and 20 respectively, which is the same in both the pictorial reporting and the household survey.

If we add seasonality into this, the picture is similar if all the surveying was done in one period, but more complicated if the household survey was spread out across different months, as I suspect may have been the case. For instance, if T1 is the low sickness period and T2 the high, then the household survey estimates would have to be appropriately weighted by the timing of the survey. Again, the best comparison would be a period-by-period comparison. So, for instance, for households surveyed in T1, I would suggest comparing the number of households who did not seek any care/report any morbidity in the one month preceding using the household survey and the pictorial diary, and similarly for T2.

In sum, I would much prefer using two specific measures:
1. The percentage of households who did not report any morbidity in the one month preceding the survey
2. The percentage of households who did not report any morbidity in the three months preceding the survey

The same measures can then be computed for health utilization. To adjust for seasonality, I would pick the appropriate months correctly. For instance, if the pictorial diary was conducted from January to June, and some households were surveyed in May, others in June, then for the May households I would compare the pictorial diary only for May and for March, April and May. I am assuming that no households were surveyed in July (that is, after the final collection of the pictorial diary) since this will constitute missing data from one source.

II. Reporting Results (Table 2): Using these two measures will make Table 2 easier to understand. Currently, Table 2 reports “Clinic Visits per year” and “Any clinic visit” for the pictorial diary and “Any clinic visit past” (which is not properly defined) and “Any clinic visit past year”, so that the two appear not to be directly comparable. I understand fully that the results yield different results, with the pictorial diary suggesting that lower user-fees led to greater utilization and the household survey method showing no result—that’s fine. At the same time, if this can be shown on precisely the same measure, it would make the point that the measurement error is not separable from the treatment effect a much stronger point. I would have preferred a full-blown regression specification with child fixed-effects, where the randomization is interacted with the data collection method to look at all effects (and the statistical significance) in a single regression framework. This can be done with and without additional adjustments.

III. Discussion and other literature:
1. One of the reasons for the difference could be that the treatment effects and the bias in measurement are non-linear, which would break the required separability of the measurement error from the treatment effect. For instance, if under-reporting increases with higher morbidity at an increasing rate, then declines in morbidity alone could generate different coefficients in the treatment effect on reported morbidity and health seeking. This is what was found in Das
and others (JDE), and the fact that the gradient of utilization with wealth was negative was one of the main findings there, leading the authors to suggest that the belief that the rich use health care more than the poor may have been directly generated by the way that data were collected.


3. In the introduction, do the authors want to suggest that there has been recent work on recall, but less on different ways of data collection, noting that their method does not allow recall effects to be separated from data collection methods?

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

I declare that I have no competing interests