Note: this pre-analysis plan is Phase 1 of two phases of this project. In this Phase 1, we perform an analysis only for one country—Liberia. After having run the code for this single county and determined any necessary adjustments, we intend to submit a second pre-analysis plan for analysis of around 45 countries in Phase 2.

Scope

1. Sample selection and geographic scope of analysis
    a. We first identified all Demographic and Health Surveys (DHS) that asked about fever in children, including the subset of surveys that included tests for malaria.
    b. We limited the study to surveys fielded between 2001 and 2014 (the period for which temperature and precipitation data is available)
    c. We choose to focus only on South and Southeast Asia, Sub-Saharan Africa, and Latin America/Caribbean. This includes more than 100 DHS surveys from more than 45 countries that asked about fever in children.
    d. Of these more than 30 surveys included tests for malaria.
    e. From this set of countries, we randomly selected a country that 1) had data for both fever and malaria; 2) had more than one year of data and 3) is in Africa. We selected Liberia.
    f. Within Liberia, as for the future full set of surveys, we will focus only on rural areas, as identified by the DHS data, and omit urban areas
    g. As stated above, this pre-analysis plan is only for Liberia. A second pre-analysis plan will follow for the full set of countries.

Data

2. Data sources – dependent and independent variables

    a. Primary dependent variable: Malaria in children under age 5 at the time of the survey (Source: DHS). DHS employs two malaria blood tests: rapid test and microscopy. We will use the rapid test as the primary dependent variable, as it was used in more countries than the microscopy test (37 vs. 32, from the earliest malaria test in 2006 through the latest available in 2015). We will follow the definitions used by the DHS.
    b. Alternative dependent variable I: microscopy test.
    c. Alternate dependent variable II: Fever in children under age 5 in the two weeks preceding the survey (Source: DHS). We will follow the definitions used by the DHS. The idea is to run all tests on all three dependent variables
    d. Independent variables (treatment)
        i. Deforestation during the year of the survey, as test of “land-cover change” hypothesis (Source: updated version of Busch and Engelmann 2015, who classified annual 30 m Landsat-derived tree-cover loss data (Hansen et al 2013/GFW) into forest or non-forest using a tree-cover threshold of 25% for 2001-2012, using same methods but adding 2013-2014).
        ii. Forest cover, as test of “land-cover” hypothesis (Source: Busch and Engelmann 2015, who classified 30 m Landsat-derived tree-cover data (Hansen et al 2013) into forest or non-forest using a tree-cover threshold of 25%). Forest cover during the

year of te survey is inferred by subtracting previous years' forest loss from forest cover in year 2000.

e. Weights. We will use the sampling weights supplied by the DHS.

3. Summary statistics:
   a. Present table of summary stats of dependent variables, independent variables, included control variables, listing all variables.

4. Superficial analyses
   a. Testing for a first-order correlation between the key outcome variables (malaria rapid test; malaria microscopy; fever;) with each other and with the main independent variable (deforestation)
      i. For each of the nine pairs we report correlation coefficient and significance (p-value)
   b. Heatmaps, to visually inspect for consistent relationship between forest cover/forest cover change and malaria:
      i. Malaria rapid test % (color) vs forest cover (x-axis, 25 increments from 0-100%) and deforestation (y-axis, 25 increments from 0 to 90th percentile)
      ii. Malaria microscopy % (color) vs forest cover (x-axis, 25 increments from 0-100%) and deforestation (y-axis, 25 increments from 0 to 90th percentile)
      iii. Fever % (color) vs forest cover (x-axis, 25 increments from 0-100%) and deforestation (y-axes, 25 increments from 0 to 90th percentile)

5. Control variables
   a. There are limitations to inferences that can be drawn from first-order correlations and heatmaps.
   b. We are concerned that third-factors could be correlated with higher or lower deforestation AND have an effect on malaria. Thus, we want to control for observable factors that have an effect on malaria.
   c. But, what to control for? DHS reports hundreds of questions from which to draw possible covariates, and then there are other data sets too. So there are a nearly infinite number of possible permutations of variables to include, and if we ran enough tests we could likely find at least some showing a positive or negative association, even in the absence of a "true" relationship (Olken, 2015).
   d. We choose variables for which there is a strong theoretic basis in the literature to have a *direct* effect on higher or lower malaria. We do not include variables that might have an *indirect* effect. Additionally, we seek to use variables that are universally available across surveys, and have a uniform interpretation across countries and time periods
   e. We specify in a pre-analysis plan which control variables we will include:
      i. Theorized proximate causes of malaria assumed to be unaffected by deforestation:
         a. Temperature (in Celsius) during the month of the survey- use cardinal values and cardinal values squared, following inverted-U-shaped relationship from literature (Beck-Johnson et al 2013, Mordecai et al 2013)
         b. Precipitation (in mm) during the month of the survey - use cardinal values and cardinal values squared, following inverted-U-shaped relationship from literature (Parham and Michel 2010)
         c. Child age – use dummy variables for each year relative to Age<1

ii. Theorized proximate causes of malaria potentially correlated with socioeconomic conditions associated with deforestation
    a. Housing quality index, constructed by summing floor type (0 for unfinished eg mud, dirt; 1 for wood, cement, etc), and wall type (0 for unfinished eg mud; 1 for finished eg wood, cement), and roof type (0 for unfinished eg thatch; 1 for finished=metal, tile, etc;)
        a. See Table 1 below for codes of individual floor, wall, and roof type
    b. Exposure to standing water: Main water source (0 if open vs. 1 if pumped/piped)
    c. Access to health care: Our preliminary intention is to use as a proxy variable a binary indicator of 1 if the child delivered in a facility and 0 otherwise. We consider this a good indicator because birth has been universally experienced by children under 5 and birth in a facility is assumed to always indicates better health.
        a. Alternative proxy indicators such as 'child has received other vaccinations', are potentially less useful because recommended vaccines vary by country, some children may be too young to have received vaccines, and there is probably a selection effect to vaccination campaigns, meaning that having been vaccinated might either mean good health OR high-risk .
        b. However, malaria indicator surveys such as those used in Liberia do not ask about place of delivery. We will investigate whether there is another acceptable indicator of access to health care that is universally asked; if not we will use this in spite of not having it for all surveys.

iii. Avoiding behavior.
    a. Bed net – 1 if "all children" or "some children" slept under a bed net last night; 0 otherwise
    b. Note: we are estimating the effect of deforestation at current levels of avoiding behavior. To the extent that deforestation increases malaria which increases avoiding behavior which dampens the increase in malaria, our estimates of the effect of deforestation on malaria are too small—that is, they have subtracted the effect of increased avoiding behavior component

f. We considered but chose not to include the following variables because they are not proximate, i.e. *direct*, causes of malaria:
    i. Wealth
    ii. Education
    iii. Remoteness
    iv. Concurrent forest-protection policy, ie. % of cell that is designated as protected

Hypothesis testing

6. Main multi-variate analysis (pooled cross-section with child-level observations)
    i. What specification to use? As with included variables, we pre-specify in order to avoid the potential to p-hack by applying log terms, fixed effects, functional forms, etc.

a. Unit of observation:
    a. survey result for child under 5
b. Functional form:
    a. Our preferred primary function form is logit, since our values are 0/1
    b. Additionally, we will consider the use of linear probability models (OLS). Preferably, we will present OLS results only as a robustness check. However, in case it is necessary to accommodate more complex features, e.g., fixed effects, clustering and sampling structure, that can't be handled by logit in Stata, then we will use OLS as our primary functional form with logit results presented as a robustness check.
c. Modification of variables, if any
    a. All variables are entered as cardinal values, that is, implying a logistic relationship with the probability of malaria, with the exception of:
        i. Forest cover, which is entered as forest cover+forest cover squared to test for hypothesized highest effects at intermediate values
        ii. Temperature and precipitation, which use cardinal values and cardinal values squared, following inverted-U-shaped relationship from literature (Beck-Johnson et al 2013, Mordecai et al 2013; Parham and Michael 2010)
        iii. Child age, which is entered in age classes
d. We cluster standard errors at the level of the grid cell because the exposure (forest cover; forest cover change) is common to all children in a grid cell.
    a. We will also assess clustering on the DHS' primary sampling unit, as a robustness check.
e. We include survey fixed effects

7. Secondary multi-variate analysis: Imbalanced panel with cell-level observations
    a. In Liberia we are fortunate enough to have multiple surveys. If it turns out that there are repeat measurements across years (i.e., subsequent DHS visited the same enumeration areas or at least areas within the same cell), which this would enable a panel regression using cell-specific fixed effects. This lets us control for the effects of spatially variant, time-invariant *unobservables* as well as *observables.* That is, we can test whether a change in deforestation leads to a change in malaria, within a cell.
        a. Aside from the addition of cell-level fixed effects, all other features of the specification are the same as above
        b. Note: in the panel analysis, we can only test the land cover change hypothesis (increased deforestation has increased malaria) not the land cover hypothesis (intermediate levels of forest cover have highest malaria)

8. Supplementary ex ante hypotheses

a. Disaggregations
    i. Geographic variation: Disease ecology might be very different in different places – our ex ante hypothesis based on the literature is that deforestation in African and Latin American countries will have a larger effect on malaria than deforestation in Asian countries.
        a. We do not test this in the Liberia-only analysis but intend to in the full analysis.
    ii. Earlier vs later forest transition: We will add as an independent variable the interaction term deforestation*forest cover. Our ex ante hypothesis is that this term will be positive, ie. deforestation at higher forest cover will have a larger effect on malaria than at lower forest cover
    iii. Smaller vs larger cuts. We will add deforestation-squared as a dependent variable. Our ex ante hypothesis is that deforestation-squared will be negative, meaning that the marginal effect of a hectare of deforestation on malaria will diminish as cuts increase in size
b. Test of mediating effects
    i. Are housing quality, health care access, and water source mediating factors through which the effect of deforestation operates? We test for this using the methods of Keele et al 2015.
        a. First, we regress the housing quality index as a dependent variable on deforestation as the independent variable, including the other control variables from the standard model (temperature, precipitation, child age, bed net usage)
        b. We also regress health care access, as proxied for with children born in a clinic as a dependent variable, on deforestation as the independent variable, including the other control variables (temperature, precipitation, child age, bed net usage)
        c. We also regress water source as a dependent variable, on deforestation as the independent variable, including the other control variables (temperature, precipitation, child age, bed net usage)
        d. Then, for each of the three mediating factors, we calculate their predicted value:
            a. with observed levels of deforestation, and
            b. without deforestation (counterfactual scenario: deforestation=0)
        e. Then, using the predictive model of malaria prevalence, calculate two predicted levels of malaria prevalence:
            a. with observed levels of deforestation and observed levels of housing quality, health care access, and water source
            b. with observed levels of deforestation but with the levels of housing quality, health care access, and water source predicted in the counterfactual scenario of no deforestation
        f. The difference between the two predicted levels in d is the effect of deforestation on malaria via housing, health care, and water source.
        g. We divide the level in f by the total effect of deforestation on malaria to calculate % of the effect of deforestation on malaria attributable to those three mediating factors

ii. Does land-cover effect mediate the effect of deforestation? We can't test for this using these methods because it violates the "sequential ignorability" assumption (Imai, Keele, and Yamomoto 2010) if land cover influences deforestation, as it likely does per Busch and Engelmann 2015. So we will not try to isolate the effect of deforestation via land cover.

c. Lagged deforestation: Replace deforestation over the last year with total deforestation over the last 7 years, following Singer and de Castro suggestion that 'frontier malaria' effect lasts 6-8 years. Only for the subset of surveys conducted 2007 or later – ex ante hypothesis: effect of last 7 years of deforestation is >1 but <7 times as large as effect of last 1 year of deforestation
d. Other disaggregations and sensitivities are reserved for exploratory analysis to generate hypotheses for future testing, rather than testing ex ante hypotheses

Methods: Total effect of deforestation on malaria

9. Hypothetically, how much lower would malaria have been in a world with no deforestation?
    a. N=all the children surveyed for malaria
    b. For each surveyed child n=1:N we use the baseline predictive model (i.e. at observed levels of deforestation) to predict the probability the child tests positive for malaria
        ii. The sum of these probabilities should be very near the number of children who actually tested positive for malaria
    c. We then predict the probability that the child would have tested positive for malaria in a counterfactual world with 10% less/50% less/100% less deforestation
    d. We divide the sum of predicted children testing positive for malaria in the counterfactual 10% less/50% less/100% less deforestation scenarios by the sum of predicted children testing positive for malaria in the baseline predicted scenario to obtain a statistic of how much lower malaria would have been with 10% less/50% less/100% less deforestation
    e. Note: this assumes no difference in avoiding behavior. To the extent that less deforestation would lead to less malaria and less avoiding behavior, our estimates would be too small.

Methods: Marginal cost effectiveness analysis

f. How cost-effective ($/DALY) is forest conservation as an anti-malarial intervention relative to other anti-malarial interventions?
    a. We want to calculate $/DALY averted from reducing deforestation, to compare with $/DALY averted from other anti-malarial interventions e.g. compared to White et al, Malaria Journal, 2011:
        i. $24/DALY for intermittent preventive treatment (IPT)
        ii. $27/DALY for insecticide-treated nets (ITN)
        iii. $143/DALY for indoor residual spraying (IRS)
        iv. Note: We need to inflate White's 2011 values to 2014$ values for comparison to Busch and Engelmann dollar values
        v. Note that these costs may not be static due to e.g. increasing drug resistance over time

b. This will be crude, as it combines data from many disparate sources with much parameter uncertainty.
c. We use a simulation to produce a density function of values, as follows (shown in Table 2 below).
    1. N=all the children surveyed for malaria
    2. For each n=1:N we "add" $100/ha/yr of forest conservation to the cell-year and estimate the resulting change in deforestation (taken from Busch and Engelmann, 2015)
        a. Alternatively, we begin with hectares rather than $ and estimate the resulting change in deforestation
    3. We then calculate the additional probability that a surveyed child within that cell-year has malaria by calculating the difference in predicted malaria between the actual deforestation and the marginally lower deforestation, using the pooled malaria regression with covariates, based on the combined land cover change effect and land cover effect.
    4. We then multiply by the "sampling fraction" f for each survey, which is the sample size (number of completed interviews) divided by the target population size (DHS Sampling Manual Sept. 2012, p.8). This works because all DHS surveys are selected so that the sample is representative of a national target population, e.g. all women age 15-49 and children under five years of age living in residential households. Stratification of sampling means the rural samples will be representative of the rural portion of the target population. We multiply malaria cases by DALY per malaria case, to obtain DALY/$
    5. Additionally, we run the regression a single time using the sample weights and "adding" $100/ha/yr of forest conservation to all cell-years. This generates a single central estimate rather than a full density function.
    6. Note: reducing deforestation has other benefits besides just reducing malaria incidence, whereas many medical interventions (treatment, bed nets) might only deal with malaria. If we were able to account for these non-malaria benefits, then the total social value of reducing deforestation relative to other interventions would be even higher.
    7. Note: surveys for malaria likely occurred in places with higher-than-average malaria areas, by design. Specifically, "[u]nlike the DHS, which is carried out at various times during the year, the MIS [Malaria Indicator Surveys] is usually timed to correspond with the high malaria transmission season. This is essential if the MIS includes biomarker testing for malaria." (http://dhsprogram.com/what-we-do/survey-

types/mis.cfm) Thus if we were to extrapolate to other areas beyond our survey, the DALY/$ would likely be lower.

Table 1

	Coded as unfinished (0)	Coded as finished (1)
Floor	adobe bamboo clay dung earth mud palm sand wood planks	brick carpet cement ceramic concrete linoleum parquet polished wood tile vinyl
Wall	adobe bamboo bark cane earth grass mud palm sticks thatch wood wood planks	bricks cement concrete metal shingles stone
Roof	bamboo cana canvas cardboard carton estera grass mud mud bricks natte palm plastic tarp sod straw thatch tin cans wood wood planks	asbestos concrete finished wood iron metal sheets shingles tiles tin zinc

Table 2

A1. Dollars -> deforestation					
Operation	Value	Assumed value	Scale of value	Uncertainty range	Source
	$\frac{\%\ decrease\ in\ deforestation}{dollars\ (\$/ha/yr)}$	Decrease at $100/ha/yr Africa: 1.60% Asia: 2.42% Latin America: 0.98%	Continental	none	Busch and Engelmann (2015)
multiplied by…	$\frac{percentage\ point\ reduction\ in\ \ deforestation}{\%\ decrease\ in\ deforestation}$	Average continent-level deforestation	Continental	none	Data Note: Continent-level rather than cell-year-level deforestation to avoid having lots of zeros
Yields…	$\frac{percentage\ point\ reduction\ in\ deforestation}{dollars\ (\$/ha/yr)}$				

A2. Hectares -> Deforestation					
	$\frac{percentage\ point\ reduction\ in\ \ deforestation}{reduced\ hectare\ of\ \ deforestation}$	Based on data		none	Data

B. Deforestation -> DALY					
Operation	Value	Assumed value	Scale of value	Range	Source
Mulitiplied by…	$\frac{reduced\ pr.a\ child\ has\ malaria\ in\ blood\ when\ sampled\ (\%)}{percentage\ point\ reduction\ of\ \ deforestation\ (\%)}$	Predicted conditional value (land cover effect+land cover change effect)	Cell-year level	Drawn from regression	Model
multiplied by…	$\frac{cases\ of\ malaria\ per\ child\ in\ a\ year}{pr.a\ child\ has\ malaria\ in\ blood\ when\ sampled\ (\%)}$	X? (365.24 days in year/how many days malaria marker lasts in blood?)	Universal	If there's a range, draw from linear distribution between high-low	Find number of days malaria lasts in blood from DHS documentation? Are samples disproportionately taken during high-malaria season? If so, need to adjust for this.
multiplied by…	$inverse\ sampling\ fraction, f^{-1} = \frac{target\ population\ size, N}{sample\ population\ size, n}$	Obtained from DHS survey documentation	Nationwide for sample year	?	DHS survey
Yields…	$\frac{\frac{cases\ of\ child\ malaria\ in\ population}{dollars\ (\$/ha/yr)}}{\frac{cases\ of\ child\ malaria\ in\ population}{additional\ hectare\ of\ deforestation}}$ or				
multiplied by…	$\frac{DALY}{case\ of\ malaria}$	4.5	Universal		(Akhaven, 1999)

		If multiple estimates exist, take their average			
Yields…	$\frac{DALY}{dollars\ (100\$/ha/yr)}$ or $\frac{DALY}{reduced\ hectare\ of\ deforestation}$				