This metadata file describes seven datasets and three R script files to repeat the analyses for the paper, “Improving rural health care reduces illegal logging and conserves carbon in a tropical forest”, published in *Proceedings of the National Academy of Sciences*. We conducted all analyses using R version 4.0.2 and RStudio Version 1.2.5042.

**Authors:** Isabel J. Jones1\*, Andrew J. MacDonald2,3,4, Skylar R. Hopkins5,6, Andrea J. Lund7, Zac Yung-Chun Liu1, Nurul Ihsan Fawzi8, Mahardika Putra Purba8, Katie Fankhauser9, Andrew J. Chamberlin1, Monica Nirmala8, Arthur G. Blundell10, Ashley Emerson11, Jonathan Jennings11, Lynne Gaffikin12, Michele Barry13, David Lopez-Carr14, Kinari Webb11, Giulio A. De Leo1,15, Susanne H. Sokolow1,16

**Affiliations:**

1 Hopkins Marine Station, Department of Biology, Stanford University, Pacific Grove, CA, USA

2 Department of Biology, Stanford University, Stanford, CA, USA

3 Earth Research Institute, University of California, Santa Barbara, CA, USA

4 Bren School of Environmental Science & Management, University of California, Santa Barbara, CA, USA

5 National Center for Ecological Analysis and Synthesis, Santa Barbara, CA, USA

6 Department of Applied Ecology, North Carolina State University, Raleigh, North Carolina

7 Emmett Interdisciplinary Program in Environment and Resources, Stanford University, Stanford, CA, USA

8 Alam Sehat Lestari, Sukadana, West Kalimantan, Indonesia

9 Oregon Health and Science University, Portland, OR, USA

10 Natural Capital Advisors, New Orleans, LA

11 Health In Harmony, Portland, OR, USA

12 Department of Obstetrics and Gynecology & Center for Innovation In Global Health, Stanford University, Stanford, CA, USA

13 Center for Innovation in Global Health, Stanford University, Stanford, CA, USA

14 Department of Geography, University of California, Santa Barbara, CA, USA

15 Woods Institute for the Environment, Stanford University, Stanford, CA, USA

16 Marine Science Institute, University of California, Santa Barbara, CA, USA

**STUDY DESIGN:**

**Intervention impact on forest change**

We used a synthetic controls analysis to test whether forest loss rates inside the focal national park declined after the intervention started (comparing the time periods 2001-2007, and 2008-2018), as compared to a synthetic control composed of all national parks across Indonesia. We ran the analysis using three sets of control groups, corresponding to the three datasets provided in the folder, “Synthetic\_controls\_analysis”. One dataset includes all 52 IUCN Category II National Parks in Indonesia as potential controls. A second dataset limits potential controls to the 44 non-marine national parks. The third dataset further limits potential controls to the 32 non-marine national parks established before 2001; results from the synthetic controls analysis using this most restrictive control group was used to communicate findings and estimate the total aboveground carbon conserved and its potential monetary value.

**Intervention impact on health care access and usage, and clinic diagnosis trends**

Health care access rates and frequency of visits between 2008 and 2018 was compared for Memoranda of Understanding (MOU)-signing patient populations to non-MOU-signing patient populations. MOU status indicates participation in the intervention. Model-estimated access and usage rates, and comparisons between the two patient populations were estimated using generalized linear mixed effects models that controlled for the travel distance required for a patient to reach the clinic.

We assessed time trends for clinic diagnoses for 11 disease categories to understand how disease trends in MOU-signing and non-MOU-signing patient populations changed over time. The disease categories assessed include: childhood-cluster diseases, chronic obstructive pulmonary disease (COPD), dental disease, diabetes, diarrheal diseases, heart disease, liver disease, lower respiratory infections, upper respiratory infections, malaria, malnutrition, neglected tropical diseases (NTDs), trauma, and tuberculosis. We used generalized linear mixed effects models to compare the proportion of patients receiving a certain diagnosis from when the clinic started (2008-2009) to the last two years of clinic data considered (2017-2018).

**Dose-response of forest change to the intervention in the focal park**

We tested whether there was any evidence of a dose-response relationship withinthe focal national park, among villages (‘dusun’) with varying levels of engagement with the intervention programs (including use of the health clinic, and other periodic programs) and forest change rates. Variation in engagement across the participating villages intersecting GPNP allowed us to test for evidence of a dose-response of intervention effort on deforestation within different access areas nearest each village around the park. Village-level access areas inside the focal park were determined by a local team that mapped the parts of each village that extended into GPNP and represented that village’s typical access area for illegal logging. We used a k-means clustering algorithm to bin village-level engagement with the intervention into “low”, “medium”, and “high” categories. Then, we used generalized linear mixed effects models to (i) estimate change in average annual forest loss rates over time, from before the intervention (2002 to 2006) to forest loss rates during the first 5 years (2007-2012) to the most recent 5 years (2013-2017), and (ii) estimate the probability that, in the last 5-year period of the intervention as compared to the 5-year period before the intervention, any 30 m2 forested pixel in a medium- or highly-engaged village’s access area within GPNP was lost, compared to the probability that any 30m2 forested pixel was lost in a low-engaged village’s access area. For the latter, we controlled for: proportion of forest lost within village boundaries outside of the park (as a proxy for outside logging pressure), average slope, and elevation of pixels inside the park, average distance of pixels to the nearest river, road, and park boundary.

**R Script Files**

**Folder: “Synthetic\_controls\_analysis”**

**GPNP\_syntheticControls.R**

This R script loads three data files:

* 52 IUCN Category II National Parks of Indonesia: **Synthetic\_controls\_allParks.csv**
* 44 non-marine IUCN Category II National Parks of Indonesia: **Synthetic\_controls\_noMarineParks.csv**
* 32 non-marine IUCN Category II National Parks of Indonesia established before 2001: **Synthetic\_controls\_noMarineParksOld.csv**

Using these data, the script runs the synthetic controls analysis evaluating the intervention’s impact on forest change rates inside the focal national park, Gunung Palung National Park. The synthetic controls analysis was run using the ‘microsynth’ package in R (https://github.com/ssdavenport/microsynth). The analysis was run using the three data files listed above, which differ only in the group of national parks included for analysis. Results from the synthetic control analysis using the 32 non-marine IUCN Category II National Parks of Indonesia established before 2001 as controls were used for communication and visualization of the intervention impact, and for estimating forest carbon conserved inside the national park and its potential monetary value.

**Folder: “Clinic\_data\_analysis”**

**usage\_&\_diagnoses.R**

This R script loads three data files:

* Data on the proportion of a district’s population that accessed the clinic at least once, district distance to the clinic, and MOU status: **ClinicPopData.csv**
* Data on the frequency that each unique patient accessed the clinic, district distance to the clinic, and MOU status: **ClinicPatientData.csv**
* Data on the number of patients diagnosed with a particular disease category (group of specific International Classification of Diseases, Revision 10 (ICD10) codes) per district, per year. Data also includes district distance to the clinic and MOU status: **ClinicDiagnosisData.csv**

Using these data, the script runs three sets of analyses: (i) statistical test (poisson mixed effects model with a population offset) on the MOU effect on the proportion of a district’s population accessing clinic at least once, controlling for each district's distance to the clinic; (ii) statistical test (negative binomial mixed effects model) on the effect of MOU on the frequency that a unique patient visited the clinic, controlling for patient's distance to the clinic; and, (ii) statistical test (binomial mixed effects model) on the change in diagnosis trends over time (first two years vs. last two years), for MOU-signing patients and non-MOU-signing patients, for ICD10 diagnosis categories of interest.

**Folder: “Forest\_loss\_dose\_response\_analysis”**

**Forest\_loss\_dose\_response\_analysis.R**

This R script loads one data file:

* Data on forest change rates in village areas bordering the focal national park, both within and outside the national park, and estimates of each village’s engagement with the interveintion over time: **dose\_response\_data.csv**

Using these data, the script runs two analyses: (i) a linear regression to estimate change in average annual forest loss rates over time, from before the intervention (2002 to 2006) to during the first 5 years (2007-2012) of the intervention to the most recent 5 years (2013-2017) of the intervention, separately for low-, medium-, and highly-engaged villages, and (ii) estimate the probability that, in the last 5-year period of the intervention as compared to the 5-year period before the intervention, any 30 m2 forested pixel in a medium- or highly-engaged village’s access area within GPNP was lost, compared to the probability that any 30m2 forested pixel was lost in a low-engaged village’s access area. For the latter, we controlled for: proportion of forest lost within village boundaries outside of the park (as a proxy for outside logging pressure), average slope, and elevation of pixels inside the park, average distance of pixels to the nearest river, road, and park boundary.

**Data Files**

**Folder: “Synthetic\_controls\_analysis”**

The “GPNP\_syntheticControls.R” script loads three data files with identical column headers, different only in the number of potential IUCN Category II National Parks included for analysis:

*Overview:* Each CSV file contains cleaned forest change data for the focal park and all potential control parks included in the synthetic controls analysis. Forest change data was obtained from the publicly available Hansen Global Forest Change dataset (version 1.6, 2000-2018, 30m pixel resolution; <http://earthenginepartners.appspot.com/science-2013-global-forest>). Each row in this dataset is one district intersecting the focal or control park, during one year.

*Description of columns:*

*DESA\_CODE*: Unique identifier for each ‘desa’ (termed ‘district’ in the manuscript) administrative unit considered, at which level forest change data and other covariate data (i.e., population) was collected and the synthetic controls analysis was run. All desa in the dataset either intersect the focal park or the IUCN Category II National Parks included as controls.

*Year*: The year for which forest change data and other covariate data is summarized.

*Forest\_Lost\_In*: The area of forest lost in a desa area *inside* the national park that it overlaps.

*Total\_Forest\_In*: The total forested area of a desa *inside* the national park that it overlaps, representing the total forested area available to log.

*Desa\_areaKm\_In:* The total area (forested or not) of a desa *inside* the national park that it overlaps.

*Fires\_In*: The total area of a desa *inside* the national park that it overlaps affected by forest fire. Forest fire data is publicly available through the MODIS Burned Area Monthly Global data product (500m), which provides the burn status of each 500m pixel at a monthly resolution (http://modis-fire.umd.edu/).

*Population\_In*: Population density in the desa area *inside* the national park it overlaps. Population density estimates were extracted from WorldPop (www.worldpop.org) by desa and year from 2000-2018.

*did*: Binary (0,1) indicator variable describing intervention status, taking a value of ‘1’ for desa in the focal park, post-intervention, and a value of ‘0’ for the control desa in all years, as well as the intervention desa, pre-intervention.

*GIS\_M\_AREA*: Total marine area of the focal park (to capture information indicating a coast adjacent or a primarily marine park).

*GIS\_AREA*: Total area of the park.

*avg\_ParkSlope*: Average pixel slope within a park (to capture ease of logging access).

*Forest\_Lost\_Out*: The area of forest lost in a desa area *outside* the national park that it overlaps, capturing logging pressure adjacent to park boundaries.

*Total\_Forest\_Out:* The total forested area of a desa *outside* the national park that it overlaps, representing the total forested area available to log.

*Desa\_areaKm\_Out*: The total area (forested or not) of a desa *outside* the national park that it overlaps.

*Fires\_Out*: The total area of a desa *outside* the national park that it overlaps affected by forest fire. Forest fire data is publicly available through the MODIS Burned Area Monthly Global data product (500m), which provides the burn status of each 500m pixel at a monthly resolution (<http://modis-fire.umd.edu/>).

*Population\_Out*: Population density in the desa area *outside* the national park it overlaps. Population density estimates were extracted from WorldPop (www.worldpop.org) by desa and year from 2000-2018.

*STATUS\_YR\_2*: The year that the national park that the desa overlaps was established.

**Folder: “Clinic\_data\_analysis”**

**File: “ClinicPopData.csv”**

*Overview:* This data includes the proportion of a district’s population that accessed the clinic at least once, district distance to the clinic, and MOU status. It is used to test for an MOU effect on the proportion of a district’s population that accessed the clinic at least once between 2008 and 2018. Each row in this dataset is one district (‘desa’).

*Description of columns:*

*Desa:* Name of each ‘desa’ (termed ‘district’ in the manuscript) in the region of the focal park included in analysis.

*MOUYN:* Binary (0,1) variable indicating whether or not an MOU was signed between the intervention program and the desa leaders.

*GIS\_time:* Estimated travel time (minutes) between the desa and the clinic. Methods to determine average distance are described in the manuscript’s Supplementary Materials.

*Desa\_pop\_2018:* Desa population as of 2018. Estimates were extracted from WorldPop (www.worldpop.org).

*UniquePatients:* Total number of unique patients that accessed the clinic between 2008 and 2018.

**File: “ClinicPatientData.csv”**

*Overview:* This data includes the number of times that each unique patient accessed the clinic between 2008 and 2018, district distance to the clinic, and MOU status. It is used to test for an MOU effect on the frequency that each unique patient accessed the clinic. Each row in this dataset is one unique patient.

*Description of columns:*

*PatientID:* Unique identifier for each patient that visited the clinic at least one time.

*Desa:* Name of each ‘desa’ (termed ‘district’ in the manuscript) in the region of the focal park included in analysis.

*MOUYN:* Binary (0,1) variable indicating whether or not an MOU was signed between the intervention program and the desa leaders.

*GIS\_time:* Estimated travel time (minutes) between the desa and the clinic. Methods to determine average distance are described in the manuscript’s SI Appendix.

*NumMonths:* The total number of months that a unique patient accessed the clinic. This excludes multiple visits per month, which were often necessary for follow-up doctor visits, to make a payment, or to pick up prescriptions.

**File: “ClinicDiagnosisData.csv”**

*Overview:* This data includes the number of times that each unique patient accessed the clinic between 2008 and 2018, district distance to the clinic, and MOU status. It is used to test for an MOU effect on the frequency that each unique patient accessed the clinic. Each row in this dataset is one disease in one village (‘dusun’) during one time period.

*Description of columns:*

*Time:* “Before” refers to the first two years that the clinic received patients, 2008-2009. “After” refers to the last two years for which clinic records were obtained, 2017-2018.

*Desa:* Name of each ‘desa’ (termed ‘district’ in the manuscript) in the region of the focal park included in analysis.

*Dusun:* Name of the village (‘dusun’ administrative unit), nested within the district (‘desa’)

*DiseasestoPlot :* Disease categories included in analysis. Details on the ICD10 codes used to summarize disease categories are available in the manuscript’s SI Appendix.

*MOUYN:* Binary (0,1) variable indicating whether or not an MOU was signed between the intervention program and the desa leaders.

*GIS\_time:* Estimated travel time (minutes) between the desa and the clinic. Methods to determine average distance are described in the manuscript’s Supplementary Materials.

*Desa\_pop\_2018:* Desa population as of 2018. Estimates were extracted from WorldPop (www.worldpop.org).

*Disease\_cat:* General disease category, including “Infectious”, “Non-infectious”, and “Other”. See manuscript and SI Appendix for details on the ICD10 codes included in each grouping.

*NumDiagnoses:* Total number of diagnoses per time period, per dusun, per unique patient, for each disease (“DiseasestoPlot”) considered.

*TotalYearlyPatients:* Total number of patients that received any diagnosis per time period, per dusun, per unique patient, for each disease (“DiseasestoPlot”) considered.

**Folder: “Forest\_loss\_dose\_response\_analysis”**

**File: “dose\_response\_data.csv”**

*Overview:* This data includes the proportion of a district’s population that accessed the clinic at least once, district distance to the clinic, and MOU status. It is used to test for an MOU effect on the proportion of a district’s population that accessed the clinic at least once between 2008 and 2018. Each row in this dataset is one village (‘dusun’) in one time period.

*Desa:* Name of each ‘desa’ (termed ‘district’ in the manuscript) in the region of the focal park included in analysis.

*Dusun:* Name of each ‘dusun’ (termed ‘village’ in the manuscript), nested within the desa.

*Year:* The last year of the 5-year periods in which forest change data was summarized. ‘2006’ refers to the period, 2002-2006; ‘2012’ refers to the period, 2008-2012; ‘2017’ refers to the period, 2013-2017.

*EngagementTotal:* The total cumulative ‘person-hours’ of engagement in various program activities, per dusun, between 2007 and 2018.

*EngagementBinned:* The engagement category, ‘low’, ‘medium’, or ‘high’ determined assigned to each dusun based on a k-means clustering algorithm.

*LossTotal5yr\_inside:* Total number of forested pixels inside the access area within the GPNP, for each dusun in the associated 5-year time period.

*ForestTotal5yr\_inside:* Total number of forested pixels inside the access area within the GPNP, for each dusun, at the start of the associated 5-year time period. This represents the total forest available to log at the start of the 5-year period in which forest loss was measured.

*PercLossAnnual\_inside:* Percent of forested pixels lost over the associated 5-year time period, in each dusun access area inside GPNP.

*LossTotal5yr\_outside:* Total number of forested pixels in a dusun area outside GPNP, for each dusun in the associated 5-year time period. This represents logging pressure outside the park.

*ForestTotal5yr\_outside:* Total number of forested pixels in a dusun area outside GPNP, for each dusun, at the start of the associated 5-year time period. This represents the total forest available to log outside the park at the start of the 5-year period in which forest loss was measured.

*PercLossAnnual\_outside:* Percent of forested pixels lost over the associated 5-year time period, in each dusun area outside GPNP.

*Slope:* Average slope of pixels in GPNP within the access area of each dusun.

*Elevation:* Average elevation of pixels in GPNP within the access area of each dusun.

*Aspect:* Average aspect of pixels in GPNP within the access area of each dusun.

*DistanceRoad:* Average distance to the nearest road of pixels in GPNP within the access area of each dusun. This is used to approximate ease of transporting logs via road.

*DistanceRiver:* Average distance to the nearest major river of pixels in GPNP within the access area of each dusun. This is used to approximate ease of transporting logs via river.

*DistanceEdge:* Average distance to the edge of GPNP within the access area of each dusun. This is used to approximate forest accessibility.

*Population\_desa:* Desa population as of 2018. Estimates were extracted from WorldPop (www.worldpop.org).