

1 **TITLE:** The Zero-Corrected, Gravity-Model Estimator (ZERO-G): A novel method to create high-  
2 quality, continuous incidence estimates at the community-scale from passive surveillance data

3

4 Michelle V Evans<sup>1,2,3\*</sup>, Felana A Ihantamalala<sup>2,3</sup>, Mauricianot Randriamihaja<sup>1,2</sup>, Andritiana  
5 Tsirinomen'ny Aina<sup>2</sup>, Matthew H Bonds<sup>2,3</sup>, Karen E Finnegan<sup>2,3</sup>, Rado JL Rakotonanahary<sup>2,3</sup>,  
6 Mbolatiana Raza-Fanomezananahary<sup>2</sup>, Bénédicte Razafinjato<sup>2</sup>, Oméga Raobela<sup>4</sup>,  
7 Sahondraritera Herimamy Raholiarimanana<sup>4</sup>, Tiana Harimisa Randrianaivalona<sup>4</sup>, Andres  
8 Garchitorena<sup>1,2</sup>

9

10 1. MIVEGEC, Univ. Montpellier, CNRS, IRD, Montpellier, France

11 2. NGO Pivot, Ranomafana, Ifanadiana, Madagascar

12 3. Department of Global Health and Social Medicine, Blavatnik Institute at Harvard Medical  
13 School, Boston, MA, USA

14 4. National Malaria Program, Ministry of Health, Antananarivo, Madagascar

15

16 \* Corresponding Author: [mv.evans.phd@gmail.com](mailto:mv.evans.phd@gmail.com)

17

18

19 **ABSTRACT**

20 Data on population health are vital to evidence-based decision making but are rarely adequately  
21 localized or updated in continuous time. They also suffer from low ascertainment rates,  
22 particularly in rural areas where barriers to healthcare can cause infrequent touch points with  
23 the health system. Here, we demonstrate a novel statistical method to estimate the incidence of  
24 endemic diseases at the community level from passive surveillance data collected at primary  
25 health centers. The zero-corrected, gravity-based (ZERO-G) estimator explicitly models  
26 sampling intensity as a function of health facility characteristics and statistically accounts for  
27 extremely low rates of ascertainment. The result is a standardized, real-time estimate of disease  
28 incidence at a spatial resolution nearly ten times finer than typically reported by facility-based  
29 passive surveillance systems. We assessed the robustness of this method by applying it to a  
30 case study of field-collected malaria incidence rates from a rural health district in southeastern  
31 Madagascar. The ZERO-G estimator decreased geographic and financial bias in the dataset by  
32 over 90% and doubled the agreement rate between spatial patterns in malaria incidence and  
33 incidence estimates derived from prevalence surveys. The ZERO-G estimator is a promising  
34 method for adjusting passive surveillance data of common, endemic diseases, increasing the  
35 availability of continuously updated, high quality surveillance datasets at the community scale.

36

37 **Key words:** health care access; geographic bias; floating catchment area; malaria; Madagascar

## 38 INTRODUCTION

39 Health metrics are vital to public health efforts, allowing decision makers to better understand  
40 the state of population health and evaluate the impact of health interventions <sup>1,2</sup>. Many of these  
41 metrics are based on routine passive disease surveillance from facility-based health  
42 management information systems (HMIS), which record the number of disease cases received  
43 at each facility at a regular frequency. Health records are then aggregated, digitized, and  
44 transferred to the district and, eventually, national health offices <sup>3</sup>. While the exact structure  
45 differs by country, the scale of spatial aggregation of the data in an HMIS corresponds to the  
46 specific level of the health system and its corresponding health infrastructure. For example,  
47 national-level data are used by international organizations to monitor long-term, multi-country  
48 trends and inform policy; regional- and district-level surveillance data may be used by national  
49 public health offices to allocate resources within the country; and individual health facility  
50 information is used by district health offices for program management.

51 Missing from most HMIS are routine surveillance data at the scale of individual  
52 communities or villages. These data are needed for spatially targeted interventions for disease  
53 control in collaboration with community health programs, which primarily serve rural  
54 communities and play an integral role in achieving universal health coverage <sup>4,5</sup>. While rural  
55 primary care facilities typically serve over ten thousand people spread along hundreds of square  
56 kilometers, community health workers (CHWs) serve between several hundred to a few  
57 thousand individuals and their catchment is generally no bigger than 10 km<sup>2</sup>. Due to geographic  
58 barriers in particular, systemic lack of access to health facilities for large portions of the  
59 population has resulted in community health becoming a central pillar of national health  
60 strategies globally <sup>6</sup>. The lack of long-term, continuously updated surveillance datasets at the  
61 community level impedes our ability to monitor changes in disease burdens over time, locally  
62 target or evaluate the impact of community-health interventions, create outbreak detection and  
63 forecasting systems at these levels, and generally incorporate health data into decision-making  
64 processes. Given the increasing role of community programs in providing primary health care  
65 and supporting disease control efforts, the lack of routine surveillance data at this level must be  
66 remedied.

67 There are several barriers to the creation of a routine surveillance system at the  
68 community level. First, CHWs often only diagnose and treat common illnesses for children  
69 under 5 years old <sup>7</sup>, representing only a subset of the population. Second, though officially part  
70 of national health systems, community health programs are often inadequately funded,  
71 supported, and integrated <sup>8,9</sup>, with negative consequences for data completeness and quality.

72 For example, a case study in Malawi found that over 40% of community health reports  
73 contained errors when aggregation was conducted by CHWs due to lack of training and time  
74 available for reporting <sup>10</sup>. Third, the existing structure of health system reporting often means  
75 that paper reports from the community level are submitted to district officials and integrated into  
76 the electronic HMIS system with significant delays, which limits their use for disease  
77 surveillance. An alternative is the use of health facility data disaggregated at the community  
78 level, which is becoming increasingly available with the development of new technologies such  
79 as eHealth systems. However, even when data remain disaggregated, there are issues of  
80 completeness and geographic bias due to heterogeneous access to care <sup>11-13</sup>. These problems  
81 are exacerbated at fine spatial scales. For example, communities in rural areas with low access  
82 to care may be missed by routine health facility systems <sup>14</sup>, significantly under-estimating  
83 disease burdens in these already vulnerable communities. Given the current lack of high-quality  
84 data at the community level, methods are needed to account for biases in these data while  
85 retaining their spatial disaggregation.

86 At the scale of the government health district and higher, several methods have been  
87 developed to address these issues, particularly under-ascertainment of cases (Table 1).  
88 However, none of these adjustment methods result in estimates of disease incidence that are  
89 available at the spatial scale of individual communities or at a temporal frequency that allows for  
90 rapid response. Existing methods are limited primarily by the frequency and spatial resolution of  
91 external data sources, such as large-scale surveys of disease prevalence or health-seeking  
92 behaviors. For example, information on healthcare utilization rates, such as that collected via  
93 Demographic and Health Surveys, is often collected nationally at the level of the district or  
94 region, and is inappropriate for use within smaller administrative zones. Prevalence surveys  
95 offer only a snapshot of disease burden in time, and their inferences, while available at finer  
96 spatial scales, often only apply to annual estimates. In addition, both forms of survey data are  
97 resource-intensive and are rarely available at spatial or temporal scales relevant to community  
98 health programs <sup>15</sup>.

100 **Table 1.** Comparing the ZERO-G method to available methods for adjusting passive surveillance data

|   | <u>Input Data</u>                                 |                             |                          | <u>Output Estimates</u>    |                           | <u>Advantages</u>   | <u>Disadvantages</u>   |
|---|---|-----------------------------|--------------------------|----------------------------|---------------------------|---|--|
|   | <b>Data Source</b>                                | <b>Frequency</b>            | <b>Spatial Scale</b>     | <b>Temporal Resolution</b> | <b>Spatial Resolution</b> |   |  |
| <b>Standard Indirect Estimators (e.g. WHO Malaria Report)</b> | Passive surveillance data for focal disease       | Annual                      | Subnational (Regional)   | Annual                     | Regional                  | <ul style="list-style-type: none"> <li>- Straightforward adjustment method</li> <li>- Directly accounts for health-seeking behaviors</li> </ul>   | <ul style="list-style-type: none"> <li>- Only available at regional or national scales</li> <li>- Requires adequate coverage of DHS surveys</li> <li>- Limited to annual estimates</li> <li>- Not appropriate for rare diseases</li> </ul> |
|   | Survey data of health-seeking behavior (e.g. DHS) | Multi-annual                | Subnational (Regional)   |                            |                           |   |  |
| <b>Ecological Downscaling (e.g. Weiss et al. 2019)</b>        | Prevalence survey                                 | Once or Multi-Annual        | Subnational (Point data) | Annual                     | 5x5 km                    | <ul style="list-style-type: none"> <li>- Avoids bias in passive surveillance data</li> </ul>  | <ul style="list-style-type: none"> <li>- Requires environmental and socio-economic variables</li> <li>- Requires prevalence data with adequate spatial coverage</li> </ul>   |
|   | Environmental Variables (e.g. Bioclim)            | Annual to Long-term Average | 5x5 km                   |                            |                           |   |  |
|   | Socio-economic variables                          | Multi-annual to annual      | Regional                 |                            |                           |   |  |
| <b>ZERO-G Estimator</b>                                       | Passive surveillance data for focal disease       | Monthly                     | Community                | Monthly                    | Community                 | <ul style="list-style-type: none"> <li>- Relies solely on health system data commonly available to Ministries of Health</li> <li>- Provides continuous, real-time estimates of incidence</li> <li>- Corrects for missing data due to data quality issues</li> </ul> | <ul style="list-style-type: none"> <li>- Requires passive surveillance data at the community level</li> <li>- Only appropriate for diseases with regular incidence and reporting</li> </ul>  |

101

102           Here, we introduce the zero-corrected, floating catchment gravity model estimator  
103 (ZERO-G). This method accounts for under-ascertainment of cases by public health facilities,  
104 resulting in a long-term dataset of disease incidence at the scale of individual communities or  
105 villages for common diseases that are regularly reported to the health system. Compared to  
106 existing methods, the ZERO-G estimator offers several distinct advantages for use in  
107 community health surveillance programs (Table 1). Because the main input data (notification  
108 reports and all-cause consultations) are released continuously on a set frequency, ZERO-G is  
109 able to produce estimates of disease incidence that are updated in real-time and available on a  
110 time scale relevant for decision makers. Unlike existing methods, ZERO-G relies solely on data  
111 available to local stakeholders: all-cause consultation rates, the focal disease incidence, and  
112 health facility characteristics. In addition, ZERO-G explicitly accounts for extremely low  
113 ascertainment rates that result in zero cases per month, a common occurrence in rural  
114 community health catchments. Finally, it does not rely on spatial aggregation or interpolation to  
115 combine estimates of healthcare utilization rates with disease incidence data, allowing it to  
116 retain a community-level spatial resolution.

117           Building on work by Hyde et al. <sup>16</sup>, the method first calculates a sampling intensity  
118 derived from healthcare utilization data (i.e. consultation rates) using a floating catchment area  
119 model <sup>17</sup>. It then uses spatio-temporal imputation to adjust for missing cases due to low  
120 healthcare access. This zero-adjusted data and the sampling intensity estimates are finally used  
121 to create an estimate of disease incidence that is adjusted for spatio-temporal heterogeneity in  
122 access to healthcare. This target diseases for this method are common, endemic diseases that  
123 are regularly reported to health systems in areas of high healthcare access (e.g. malaria,  
124 pneumonia, diarrheal disease). ZERO-G is not appropriate for rare diseases or those where  
125 only severe cases are reported. We demonstrate the method on a simulated endemic disease  
126 and on a case-study of field-derived passive surveillance dataset of malaria in a rural health  
127 district in southeastern Madagascar. The case study is used to further validate the ZERO-G  
128 method by comparing the estimated sampling intensity and malaria incidence rates to health-  
129 care seeking behavior and malaria prevalence from a district-representative cohort.

130

### 131 **THE ZERO-G ESTIMATOR**

132 Indirect estimation methods estimate the “true” rate of disease incidence or prevalence from  
133 case data with low or uneven ascertainment rates by including information on the sampling  
134 intensity (e.g. healthcare use) in each administrative region <sup>18</sup>. ZERO-G specifically combines

135 information on the number of cases recorded by the health system with information on the  
 136 proportion of cases that are expected to be observed. In addition, it includes imputation  
 137 methods for adjusting for extremely low ascertainment rates that result in zero cases reported.  
 138 The final result is an estimation of disease incidence equal to the expected incidence if access  
 139 to healthcare was identical across space and time.

140 The ZERO-G estimation method can be summarized in a pseudo-statistical framework  
 141 consisting of three main steps (Figure 1): 1) the estimation of healthcare access via a gravity  
 142 model (Eq. 7-12), 2) the adjustment of erroneous zeroes in case notifications (Eq. 4-6) and, 3)  
 143 the conversion of healthcare access to sampling intensity via multi-objective optimization (Eq. 2-  
 144 3). The estimates of sampling intensity and zero-adjusted data are then used to estimate an  
 145 adjusted incidence rate ( $N_{it}$ ) for each administrative zone  $i$  and time period  $t$ , accounting for  
 146 imperfect detection due to differing healthcare access via an Inverse Binomial distribution (Eq.  
 147 1). The full ZERO-G estimator can be stratified across demographic classes (e.g. age, sex, etc.)  
 148 to account for demographically-dependent health-seeking behaviors. However, we limit our  
 149 notation here to one class to improve readability. Parameters and variables representing data  
 150 are further described in Table 2.

151

$$152 \quad N_{it} \sim InvBin(r^*_{it}, SI_{it}) \quad \text{(Equation 1)}$$

153

154 *Rescaling healthcare access to sampling intensity (SI) via multi-objective optimization*

$$155 \quad SI_{it} = \left( \frac{(1 - x_1) * (A_{it} - x_1)}{\max(A) - \min(A) + x_1} \right)^{x_2} \quad \text{(Equation 2)}$$

$$156 \quad \min_{x \in X} (f_1(x), f_2(x), f_3(x), f_4(x)), X \subseteq \mathbb{R}^2 \quad \text{(Equation 3)}$$

157 *Zero-adjustment framework*

$$158 \quad r^*_{it} = \begin{cases} RF(X_i, Y_i, m_t) & \text{if } r_{it} = 0 \text{ \& } \psi_{it} < 0.5 \\ r_{it} & \text{else} \end{cases} \quad \text{(Equation 4)}$$

159

$$160 \quad Z_{it} \sim Bernoulli(\psi_{it}) \quad \text{(Equation 5)}$$

$$161 \quad \text{logit}(\psi_{it}) = \beta_{z0} + \beta_{z1}m_t + \beta_{z2}A_{it} + \beta_{z3}m_tA_{it} \quad \text{(Equation 6)}$$

162

163

164 *Estimating healthcare access (A) via a gravity model*

$$165 \quad h_{it} \sim Binomial(n_{it}, A_{it}) \quad \text{(Equation 7)}$$

166

$$A_{it} = g(t) \sum_j \frac{S_{jt} f(d_{ij})^2}{C_{jt}} \quad (\text{Equation 8})$$

167

$$S_{jt} = \sum_s \beta_s v_{jst} \quad (\text{Equation 9})$$

168

$$f(d_{ij}) = e^{-\lambda d_{ij}} \quad (\text{Equation 10})$$

169

$$C_{jt} = \beta_c \sum_k n_{kt} f(d_{kj}) \quad (\text{Equation 11})$$

170

$$g(t) = e^{\beta_{g1}t + \beta_{g2} \frac{\sin(2\pi(t+\phi)/12) + 1}{2}} \quad (\text{Equation 12})$$



171 **Table 2.** Description of variables and parameters in ZERO-G method.

| Variable          | Description  | Source                |
|-------------------|--|-----------------------|
| $\bar{N}_{it}$    | The ZERO-G estimated incidence after correcting for under-ascertainment  | Estimated via Eq. 1   |
| $r_{it}^*$        | Zero-adjusted reported incidence   | Estimated via Eq. 4   |
| $RF(X)$           | Function describing random forest algorithm used in data imputation  | Estimated via Eq. 4   |
| $SI_{it}$         | The sampling intensity in zone $i$ at time $t$   | Estimated via Eq. 2   |
| $r_{it}$          | The reported incidence in zone $i$ at time $t$   | From data             |
| $X_i$             | Longitude of zone $i$  | From data             |
| $Y_i$             | Latitude of zone $i$   | From data             |
| $m_t$             | Month of year (e.g. Jan-Dec) at time $t$   | From data             |
| $\psi_{it}$       | Probability of zero reported incidence   | Estimated via Eq. 5   |
| $Z_{it}$          | Binomial variable representing if there was zero reported incidence  | From data             |
| $\beta_{z1...zn}$ | Coefficients used to estimate probability of zero reported incidence   | Estimated via Eq. 6   |
| $A_{it}$          | FCA-based healthcare access  | Estimated via Eq. 7,8 |
| $h_{it}$          | Number of non-focal disease consultations  | From data             |
| $n_{it}$          | Population count of zone   | From data             |
| $g(t)$            | Function describing temporal trend in $A_{it}$ . Specific function can be adjusted based on need                           | Estimated via. Eq. 12 |
| $S_{jt}$          | Services provided by health facility $j$ at time $t$   | Estimated via Eq. 9   |
| $C_{jt}$          | Competition at health facility $j$ at time $t$   | Estimated via. Eq. 11 |
| $v_{jst}$         | Value of health service $s$ provided at health facility $j$ at time $t$  | From data             |
| $\beta_s$         | Coefficient for health service $s$   | Estimated via Eq. 9   |
| $d_{ij}$          | Distance between zone $i$ and health facility $j$ . Can be calculated using Euclidean distance or based on actual routing. | From data             |
| $\lambda$         | Distance decay coefficient   | Estimated via Eq. 10  |
| $\beta_c$         | Scaling coefficient for competition at health facilities   | Estimated via Eq. 11  |
| $x_1, x_2$        | Scaling coefficients for SI.   | Estimated via Eq. 6   |

172 Note: Subscript  $i$  refers to zone  $i$  and subscript  $t$  refers to time  $t$ .

173

174 Healthcare access ( $A$ ) is estimated from monthly healthcare utilization rates (i.e.

175 consultation rates with the focal disease removed,  $h_{it}$ ). The relationship between the monthly

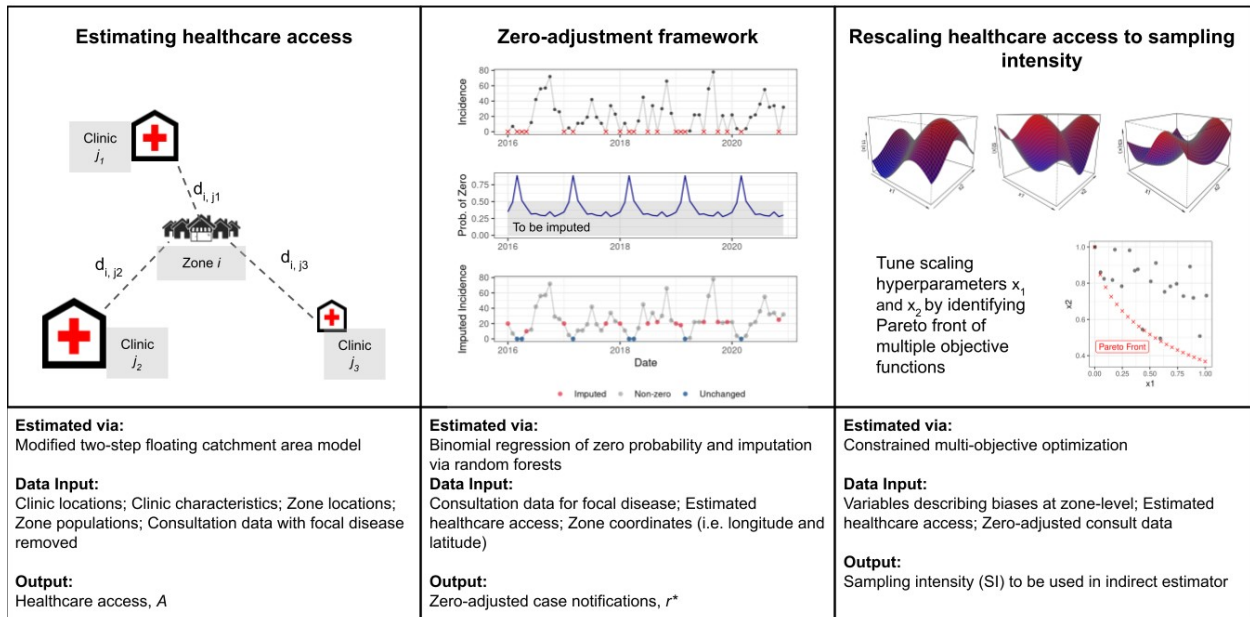
176 number of consultations and the estimated healthcare access is defined via a temporally-explicit

177 floating catchment area (FCA) model of healthcare access (Eq. 7-12)<sup>19</sup>. Based on a gravity  
178 model, FCA models consider both the quantity and spatial accessibility of services at a health  
179 center for a given population by weighting the distance to care by the availability of services  
180 provided at each health center. The effect of distance on healthcare access is described by a  
181 function  $f(d_{ij})$  that assumes exponential distance-decay, with the specific shape of the decay  
182 defined by  $\lambda$  (Eq. 10). Healthcare access in each community is then modeled via an “attractive  
183 force” to each health center and total access to care is the sum of these forces for a given zone  
184 (Eq. 8). Specifically, we use the modified two-step floating catchment area formulation of this  
185 metric, which allows for sub-optimal allocation of health resources via the inclusion of distance-  
186 weighted competition for each healthcare clinic’s resources<sup>20</sup>. We also include a term to  
187 account for temporal trends in access due to seasonal and linear trends (Eq. 12), following  
188 Garchitorena et al. 2021<sup>11</sup>.

189 Erroneous zeroes to be adjusted are a function of both the seasonality of the disease  
190 and the estimated healthcare access of each zone. They are identified by fitting a logistic  
191 regression to the binomial variable of whether zone  $i$  reported zero incidence at time  $t$  (Eq. 4-5),  
192 resulting in estimates of the probability of a zero ( $\psi_{it}$ ). The logistic regression’s explanatory  
193 variables include the month of the year of time  $t$ , estimated healthcare access for zone  $i$  at time  
194  $t$ , and the interaction between the two (Eq. 6). This logistic regression is fit to the reported case  
195 data to estimate  $\psi_{it}$ . If zero cases are reported in a month for a zone and the probability of  
196 reporting zero cases is less than 0.5, this zero is assumed to be due to reporting error (not  
197 seasonality or low access) and is defined as erroneous. Erroneous zeros are replaced via a spatio-  
198 temporal imputation process that incorporates seasonal and spatial patterns in incidence.  
199 Imputation is performed via 100 boosted regression tree models that estimate monthly incidence  
200 as a function of each zone’s longitude, latitude, and specific month of the zero-incidence  
201 occurrence, using the median of 100 imputations as the final imputed value (Eq. 4). Imputation is  
202 performed via the micemd package v 1.9.0 in R<sup>22</sup>.

203 The sampling intensity (SI) is calculated from healthcare access via a constrained multi-  
204 objective optimization routine that minimizes four objective functions (Eq. 3). The objective  
205 functions correspond to: 1) the Spearman correlation coefficient between a zone’s distance to a  
206 PHC and its average annual incidence rate (geographic bias), 2) the ratio of incidence rates in  
207 zones with reimbursement policies to those without (financial bias), 3) the number of zones with  
208 annual incidence rates over 1000 cases per 1000 population (over-correction bias), and 4) the  
209 covariance of all three biases, to reduce over-correcting one value at the expense of the others.  
210 This creates a Pareto front of non-dominated values across the four objectives. From this

211 subset, a constraint is used to limit the over-estimation of cases by constraining the results to  
 212 parameters that result in monthly incidence values where the 99% percentile falls below a  
 213 threshold equal to 1.5 times the original maximum monthly incidence value. The optimization  
 214 routine is solved using the NSGA-II genetic algorithm via the mco package in R <sup>21</sup>. The optimal  
 215 values of  $x_1$  and  $x_2$  are then used to rescale  $A_{it}$  between  $x_1$  and 1 to calculate  $SI_{it}$  (Eq. 2).  
 216  
 217



218  
 219 **Figure 1. Workflow for adjusting incidence data using the floating catchment, zero-**  
 220 **corrected (ZERO-G) estimator. Panel 1:** A depiction of the gravity-model used in the floating  
 221 catchment area model. A single zone  $i$  is represented surrounded by multiple clinics  $j$  with  
 222 differing amount of services offered, with the distance between the zone and the clinic  
 223 represented by  $d_{ij}$ . **Panel 2:** An example of the zero-adjustment step for one zone. Top row of  
 224 Panel 2: All zeroes are identified in the dataset, represented by an X. Middle row of Panel 2:  
 225 The probability of a zero is estimated via a logistic regression and those samples with a  
 226 probability below 0.5 are identified. Bottom row of Panel 3: Those zeros that occur during a  
 227 month with less than 0.5 probability of a zero are replaced via an imputation step. **Panel 3:**  
 228 Hyperparameters are tuned via multi-objective optimization across a hyper-dimensional space,  
 229 resulting in a Pareto front of non-dominated parameter values.

230  
 231  
 232

233 **CASE STUDY: MALARIA INCIDENCE IN IFANADIANA, MADAGASCAR**

234 We applied the ZERO-G estimator to malaria incidence in Ifanadiana District, Madagascar to  
 235 demonstrate its utility in regions with highly heterogeneous rates of under-ascertainment.

236 Ifanadiana is a district in the Vatovavy region of southeastern Madagascar. It has an estimated  
237 population of 183,000 people spread across 195 fokontany (smallest administrative unit  
238 comprising about 1000 people) within 15 communes. Each commune contains one primary  
239 health center level 2 (PHC2), and six of the larger communes also contain a primary health  
240 center level 1 (PHC1), which provides more basic care, for a total of 21 PHCs within the district.  
241 Beginning in 2014, the Madagascar Ministry of Public Health (MMoPH) and the non-  
242 governmental organization Pivot began a partnership to strengthen the health system,  
243 establishing Ifanadiana as a model health district. This intervention works across all levels of the  
244 health system, from community health at the household level to tertiary care at the regional  
245 hospital. At the level of the PHCs, in addition to the removal of user fees, the intervention  
246 includes a range of activities to increase PHC readiness (e.g. infrastructure, equipment,  
247 supplies and personnel), support clinical programs (e.g. maternal and child health, infectious  
248 diseases), and improve data systems. As of January 2023, a minimum package of support has  
249 been provided to all 15 PHC2s of all 15 communes, and will be expanded to a complete  
250 package at all levels of PHCs by the end of 2024. Because the progress of these health system  
251 strengthening interventions in Ifanadiana and elsewhere typically differ across PHCs and time,  
252 this requires an adjustment method that considers spatio-temporal differences in healthcare  
253 policies and interventions, such as the ZERO-G estimator.

254 As is common in sub-Saharan Africa<sup>24</sup>, the primary barriers to healthcare at PHCs in  
255 Ifanadiana are geographical and financial. The majority of the district is rural and the  
256 transportation network is primarily non-motorized; over 70% of the population lives further than  
257 an hour travel time from a PHC<sup>25</sup>. As such, geographical access to care at PHCs is highly  
258 unequal, and exhibits strong distance-decay from PHC locations<sup>11</sup>. Regarding financial  
259 barriers, 34% of the public health expenditure in Madagascar is out-of-pocket spending<sup>26</sup>, with  
260 user fees the most cited barrier to healthcare seeking across the district<sup>27</sup>. Given these known  
261 barriers, we aimed to reduce the impact of geographic and financial bias in malaria incidence  
262 rates by adjusting the data using ZERO-G.

263

#### 264 **Data Collection**

265 Monthly consultation data were collected at each PHC for the district of Ifanadiana from January  
266 2016 to December 2021. Photos were taken of handwritten registries at each PHC, and  
267 patients' residences were manually geolocated to the precision of the fokontany. The number of  
268 all-cause consultations were reported by fokontany, as well as the number of malaria cases, as  
269 confirmed by rapid detection test. Because patient ages were provided in these registries, we

270 were able to divide the number of consultations and malaria cases into three age groups for  
271 analysis: children under 5 years old, juveniles aged 5-14, and adults aged 15 and over.  
272 Ifanadiana suffers from shortages of diagnostic materials, specifically rapid-detection tests  
273 (RDTs)<sup>28</sup>, leading to unconfirmed cases of malaria. We accounted for this reduced diagnostic  
274 capacity by scaling the confirmed malaria cases by the proportion of feverish patients who were  
275 tested via an RDT at each PHC during each month (n = 536). Information on the characteristics  
276 of each clinic by month was provided by Pivot's Monitoring and Evaluation for Research and  
277 Learning team.

278         Population data came from two sources. For the 80 fokontany that receive community  
279 health program support from Pivot, we used population estimates from a Pivot-led census  
280 conducted in 2021. For the remaining 115 fokontany, population estimates came from a national  
281 census conducted in 2018 by the Madagascar National Institute of Statistics. By interpolating  
282 population values between the 2018 census and the previous 1993 census, we estimated an  
283 average annual population growth rate of 2.0%. We applied this population growth rate to both  
284 datasets to obtain each fokontany's population by year. For both datasets, we assumed 18% of  
285 the population to be under 5 years old, 28.6% of the population to be aged 5 - 14 and the  
286 remainder to be 15 years old or above, based on the average age structure of the 80 fokontany  
287 that were censused in 2021.

288         Distances between residential areas and PHCs were calculated on a high-resolution  
289 transport network created via crowd-sourced mapping through a collaboration with  
290 Humanitarian OpenStreetMap. Over 20,000 km of footpaths and 100,000 buildings within the  
291 district were mapped through a two-step validation process<sup>25</sup>, resulting in an open-source  
292 dataset on OpenStreetMap. Using this dataset, we estimated the distance between each  
293 household and each PHC within the district, and aggregated this to the scale of the fokontany to  
294 result in an average distance to each PHC for each fokontany. Three fokontany lacked accurate  
295 routing information and so were excluded from the analysis.

296         We evaluated our estimates of the SI and adjusted malaria incidence rates using  
297 external data from a longitudinal cohort survey conducted in the district of Ifanadiana (IHOPE  
298 cohort). The IHOPE cohort has conducted population-representative surveys approximately  
299 every two years from 2014-2021 using a two-stage cluster sampling scheme involving 80 spatial  
300 clusters, each containing 20 households<sup>29</sup>. We include data from 2016, 2018, and 2021 in this  
301 analysis. The IHOPE cohort is based on the internationally validated Demographic and Health  
302 Surveys and is implemented by the Madagascar National Institute of Statistics. See Miller et al.  
303<sup>29</sup> for further details on participant recruitment and study design. As part of the survey

304 questionnaire, participants were asked if they were ill in the past four weeks and, if so, if they  
305 sought care at a public PHC. This data represented self-reported health-care seeking behavior,  
306 comparable to ZERO-G estimates of sampling intensity. Malaria prevalence data was collected  
307 via rapid detection tests conducted as part of the IHOPE survey in 2021. Briefly, children under  
308 15 years old who consented to the study were tested for active malaria infection using SD One  
309 Step Malaria HRP-II(P.f) and pLDH(Pan) Antigen Rapid Tests. Those who tested positive were  
310 provided with a standard treatment of artesunate amodiaquine and paracetamol, with duration  
311 and dosage in accordance with national guidelines. In total, this resulted in 3774 samples  
312 across 80 clusters and 109 fokontany.

313

### 314 ***Applying the ZERO-G estimator***

#### 315 *Estimating healthcare access (A)*

316 We estimated the healthcare access for each fokontany and month combination in our dataset  
317 following the methods described above for each age class (children, juveniles, and adults) using  
318 non-malarial consultations at PHCs. We included five traits of the health center in our  
319 calculation of  $S_j$ :

- 320 1. whether the PHC fell within the initial Pivot service catchment,
- 321 2. if point-of-care user fees (consultation costs and medications) had been removed at that  
322 time,
- 323 3. the number of staff at the PHC during each month,
- 324 4. level of health clinic (PHC1 or PHC2, with PHC2 providing more services),
- 325 5. distance from the PHC to the District office, which provides supplies, medications, and  
326 supervision.

327 In addition, two new PHC2 were opened in the district during the study period, one in  
328 Ampasinambo in November 2016 and one in Ambiabe in April 2018, which we accounted for in  
329 our calculation of SI. Notably, ZERO-G allows for health center traits that change over time,  
330 which we used to include monthly staffing changes, the construction of new health centers, and  
331 user fee removal interventions that were implemented over the study period.

332 To reduce computational time, we set a maximum limit on the distance between a  
333 community and the PHC ( $d_{ij}$ ) at 25 km, slightly above the maximum distance of a fokontany to  
334 the nearest PHC in Ifanadiana (22.1 km). We also included an additional parameter in our  
335 estimation of  $f(d_{ij})$  to allow the shape of this relationship to differ for those fokontany within the  
336 Pivot zone of intervention and those outside of the zone of intervention, following Garchitorena  
337 et al.<sup>11</sup>

338 We estimated the number of non-malarial consultations  $h_{it}$  as a random variable with a  
339 binomial distribution with the probability equal to the healthcare access ( $A_{it}$ ) and size  $n_{it}$  equal  
340 to the population size of the fokontany (Eq. 7). Some fokontany had extremely low consultation  
341 rates and reported zero consultations for over 50% of the study period. We excluded these  
342 fokontany ( $n = 43$ ) from the model fitting exercise estimating the parameters for  $A_{it}$ , but did  
343 estimate their healthcare access from the fit model. To ensure our estimate represented the  
344 global maximum likelihood estimate (MLE), and not a local maximum, we used a two-step MLE  
345 estimation process. First, we performed a grid search via a latin hypercube sample of 1,000  
346 samples of coarse parameter space to identify the ten parameter sets with the lowest negative  
347 log-likelihood. We then performed a second MLE step using the BFGS algorithm via the optim  
348 function in the *stats* package in R <sup>30</sup>, using the parameters from the ten parameter sets with the  
349 lowest negative log-likelihood from the first step as the starting parameters. We assessed each  
350 of these ten iterations for convergence and selected the parameter set with the lowest negative  
351 log-likelihood as the optimal fit. A total of 11 parameters were estimated for each age class  
352 (Table S2.1). From the optimal parameter sets, we estimated  $A_{it}$  for each fokontany-month  
353 combination for each age-class via Eq. 8.

354

355

#### 356 *Imputing erroneous zeroes*

357 Nearly all fokontany ( $n = 189$ ) reported zero malaria cases across all ages at least once during  
358 the study period, totaling 3468 (28.4%) of fokontany-month samples. On average, fokontany  
359 reported zero malaria cases for 18.1 months out of the 66 month period, with a range of 0 - 53  
360 months reporting zeros. The ZERO-G method imputed between 6.08 to 10.00% of fokontany-  
361 month incidence values for each age class, an average of 4.75 months per fokontany (range: 0  
362 – 22).

363

364

#### 365 *Rescaling healthcare access to sampling intensity*

366 We manually set the sampling intensity (SI) to 1 for those fokontany which had an average  
367 annual healthcare utilization rate over 1 consultation per capita-year, defined as “high access  
368 fokontany” ( $n = 19$ ). The remaining fokontany’s healthcare access values were rescaled  
369 following Eq. 2 and 3 using multi-objective optimization to calculate their monthly SI values.

370

371

372

### 373 ***Evaluating Adjusted Datasets***

374 We evaluated our estimates of the SI and adjusted malaria incidence rates using external data  
375 from the IHOPE cohort. Self-reported healthcare seeking behavior was paired spatially to SI  
376 estimates by assigning a value to a fokontany if a village from the cluster was in that  
377 fokontany. These data were paired temporally by taking the average of the SI during the 6  
378 month period containing the months when the IHOPE survey was conducted in each year  
379 (January through June for 2016 and 2021 and July through December for 2018), to reduce the  
380 impact of month outliers in healthcare utilization data on SI estimates. We assessed the  
381 agreement between the two datasets by calculating the correlation between estimated SI and  
382 the proportion of residents reporting illness who attended PHCs using Clifford's modified t-test,  
383 which controls for spatial autocorrelation<sup>32</sup>. We assessed the correlation separately for each  
384 year (2016, 2018, 2021), including 109 fokontany per year.

385 We evaluated the ability of ZERO-G adjusted incidence rates to accurately represent  
386 malaria burdens by comparing adjusted incidence rates to malaria prevalence data collected via  
387 the IHOPE cohort in 2021. The two datasets were paired spatially by assigning a value to a  
388 fokontany if a village from the cluster was in that fokontany and were paired overtime by  
389 matching the month of the IHOPE survey to the month of the incidence rates. Because the  
390 relationship between prevalence and incidence is non-linear, we transformed cluster-level  
391 prevalence rates into incidence rates following a previously published model<sup>33</sup> to allow us to  
392 compare incidence rates from both datasets. However, there remain important differences  
393 between this measure of incidence and that derived from case notifications. Prevalence data  
394 may under-estimate malaria incidence as the conversion results only in symptomatic cases of  
395 malaria while case notifications may include a higher proportion of asymptomatic cases due to  
396 co-infection with a second febrile-inducing pathogen<sup>34</sup>. We compared adjusted incidence rates  
397 for children under 15 years old to prevalence rates of children under 15 years old from the  
398 IHOPE cohort for all fokontany with information in both datasets (n= 109) via Clifford's modified  
399 t-test. We also assessed the ability of the adjusted incidence data to correctly identify hot spots  
400 of malaria, defined as the quartile of fokontany with the highest prevalence values.

401

### 402 **APPLYING ZERO-G TO A SIMULATED DISEASE**

403 To demonstrate its generalization, we used the ZERO-G estimator to adjust for under-  
404 ascertainment of cases of a simulated endemic, seasonal disease. We simulated a model



405 health district containing 100 administrative zones and 8 health clinics that differed in the  
406 number of staff, whether they offered advanced services, and whether health care was  
407 subsidized. We then simulated disease dynamics for a constant background disease rate and  
408 for two additional diseases that exhibited annual seasonality for each administrative zone at a  
409 monthly frequency for five years. We modeled an individual's probability of seeking care as a  
410 random variable with probability equal to that zone's reporting rate, itself a function of its  
411 distance to a clinic and the services available at that clinic, plus a random error term (Eq. S2).  
412 To represent realistic issues in data quality, we also simulated months reporting zero cases as  
413 both a function of low reporting rates and low disease incidence and due to randomness. This  
414 resulted in a time series of "true" disease incidence and reported disease incidence for each  
415 zone over a five year period (Fig. S3). Further details on the creation of the simulated dataset  
416 are reported in the Supplemental Materials.

417         The performance of the ZERO-G method on the simulated data was evaluated by  
418 comparing the ability of the ZERO-G to reproduce the original simulated "true" data. We  
419 calculated the root mean squared error (RMSE) and correlation coefficient between the true  
420 incidence and adjusted incidence rates across patches and seasons. We compared these  
421 values to the unadjusted incidence rates to assess the improvement provided by the ZERO-G  
422 method.

423

#### 424 ***ETHICS STATEMENT***

425 Use of aggregate monthly healthcare utilization data from PHCs in Ifanadiana District for this  
426 study was authorized by the Medical Inspector of Ifanadiana. The IHOPE longitudinal survey  
427 implemented informed consent procedures approved by the Madagascar National Ethics  
428 Committee and the Madagascar Institute of Statistics. Household-level de-identified data from  
429 the IHOPE survey were provided to the authors for the current study. We recognize that all  
430 research is conducted within the surrounding socio-political context and risks reproducing  
431 existing inequalities within the research team and across research partners. We've chosen to  
432 explicitly reflect on power dynamics and equitable authorship within the context of this research  
433 project in an accompanying reflexivity statement (Supplemental Materials).

434

#### 435 **RESULTS**

436

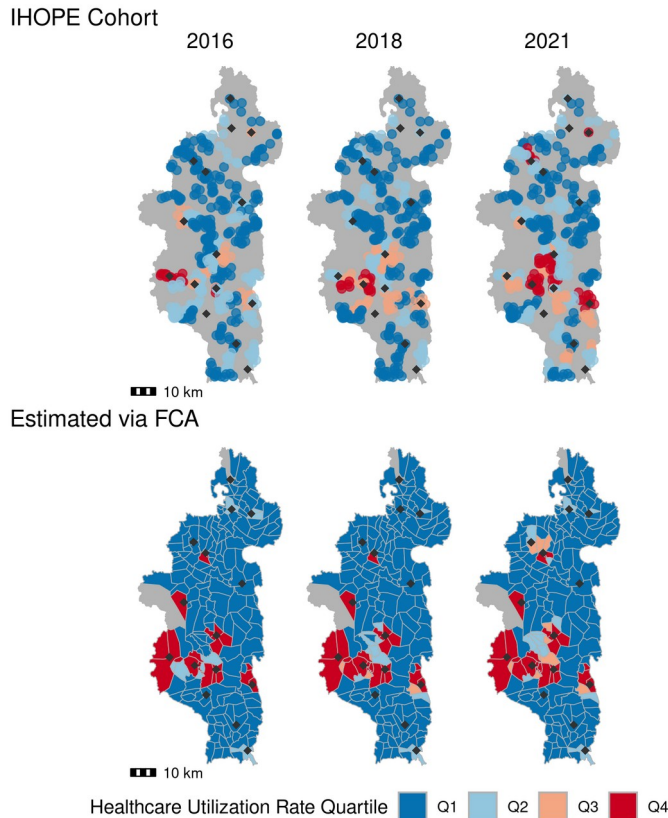
437

438 **Case Study: Malaria in Ifanadiana, Madagascar**

439 We estimated the SI by fitting a floating catchment area model to healthcare utilization data from  
440 January 2016 - December 2021 and rescaling it via multi-objective optimization. The resulting  
441 model performed well at reproducing the healthcare utilization data (under-5: Spearman's  $\rho =$   
442 0.619,  $p$ -value<0.001; juvenile: Spearman's  $\rho = 0.608$ ,  $p$ -value<0.001; adult: Spearman's  $\rho =$   
443 0.702,  $p$ -value<0.001). When averaged over all fokontany per month, it accurately represented  
444 the temporal trends in the healthcare utilization data, although this performance was dependent  
445 on age-class (under-5: Spearman's  $\rho = 0.384$ ,  $p$ -value <0.01; juvenile: Spearman's  $\rho = 0.517$ ,  $p$ -  
446 value <0.001; adult: Spearman's  $\rho = 0.578$ ,  $p$ -value < 0.001). When averaged across time to  
447 result in one average SI per fokontany, it also was able to capture spatial and fokontany-specific  
448 differences in healthcare utilization rates (under-5: Spearman's  $\rho =0.829$ ,  $p$ -value <0.001;  
449 juvenile: Spearman's  $\rho =0.806$ ,  $p$ -value <0.001; adult: Spearman's  $\rho =0.844$ ,  $p$ -value <0.001).

450 The spatial patterns in the estimated SI mirrored spatial patterns in self-reported  
451 healthcare seeking behavior from the IHOPE longitudinal survey (Fig. 2). The estimated SI and  
452 self-reported healthcare seeking rates were significantly correlated across all years (Clifford's t-  
453 test; 2016:  $\rho = 0.502$  ( $p < 0.01$ ), 2018:  $\rho = 0.644$  ( $p < 0.01$ ), 2021:  $\rho = 0.564$  ( $p < 0.01$ ), Fig  
454 S2.1). Both data sources estimate higher healthcare access at fokontany nearer the national  
455 transportation network, specifically the paved road that runs east-west through the district, and  
456 in close proximity to PHCs. In addition, the two datasets were in agreement that the majority of  
457 the district has low access to healthcare.

458



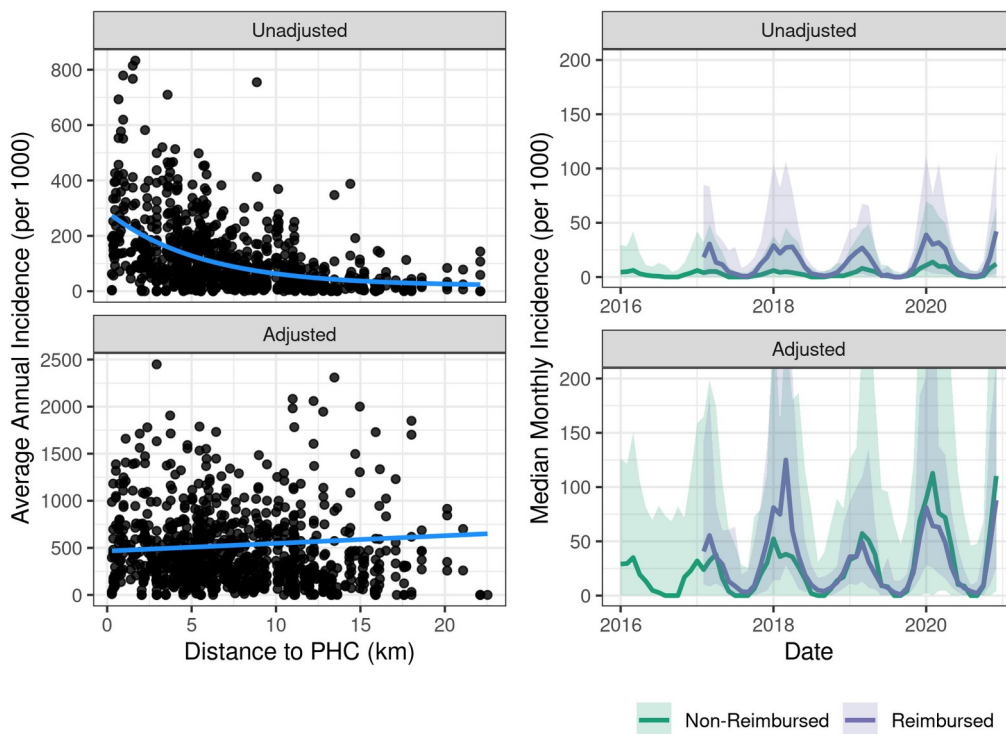
459

460 **Figure 2. The sampling intensity estimated via the gravity model and multi-objective**  
 461 **optimization (bottom row) closely approximates self-reported healthcare seeking rates**  
 462 **from the IHOPE cohort (top row).** Shading represents rates grouped into quartiles, with Q1  
 463 corresponding to the lowest healthcare utilization rate and Q4 to the highest. Diamond points  
 464 show the location of level-2 PHCs. Top row: Cluster-level healthcare seeking rates are  
 465 illustrated for each village in a cluster across the three survey years. Bottom row: The scaled  
 466 sampling intensity estimated via Eq. 2-3 & 7-12. Scatter plots of this data are shown in Fig S2.1  
 467

### 468 ***Reduction of Bias in Malaria Incidence due to Geographic and Financial Barriers to Care***

469 The unadjusted dataset showed evidence of geographic bias; average annual incidence of  
 470 malaria in a fokontany was negatively correlated with the distance from that fokontany to the  
 471 nearest PHC (Spearman's  $\rho = -0.617$ ,  $p$ -value  $< 0.001$ , Fig. 3A), showing an exponential  
 472 distance decay. The adjusted dataset, by comparison, demonstrated no relationship between  
 473 average annual incidence and distance to the nearest PHC (Spearman's  $\rho = -0.060$ ,  $p$ -value =  
 474 0.409, Fig. 3A). Fokontany whose populations attended PHCs where fees were removed for  
 475 the user (PHCs were reimbursed by Pivot) reported 2.48 times higher incidence than those that  
 476 did not benefit from the reimbursement policy in the unadjusted dataset (Fig. 3B). Applying the  
 477 ZERO-G method drastically reduced this bias; the average annual incidence in these fokontany  
 478 was 0.95 times the incidence in fokontany with cost-of-care-reimbursement (Fig. 3B). However,

479 this reduction in bias differed across years. Specifically, zones with reimbursement policies  
 480 retained a much higher incidence rate in 2018. This difference was driven primarily by high  
 481 monthly incidence (>500) in the unadjusted data due to a malaria outbreak in the north of the  
 482 district in a commune benefiting from fee-reimbursement. Because it does not aggregate or  
 483 smooth incidence data, ZERO-G retained this anomaly in incidence rates even after adjustment.  
 484 This is an advantage of ZERO-G, as it allows for the identification of epidemics or unexpected  
 485 trends in the data.  
 486



487  
 488 **Figure 3. The ZERO-G adjustment method greatly reduces geographical and financial**  
 489 **bias in malaria incidence rates.** Left: Each point represents the average annual malaria  
 490 incidence rates for a fokontany over the period of 2016-2020, with the x-axis showing the  
 491 distance to the nearest PHC. The smoothed line is the exponential (unadjusted) or linear  
 492 (adjusted) fit between average annual incidence and distance to PHC. One outlier point is  
 493 removed to aid with visualization. Right: The median monthly malaria incidence rates across  
 494 fokontany whose closest PHC does or does not offer fee reimbursement. Fee reimbursement  
 495 began in January 2017. The error ribbon represents a 90% CI. The y-axis is limited between  
 496 values of 0-200 to aid with visualization.  
 497

498 **Comparing Unadjusted and Adjusted Datasets**

499 Comparing the unadjusted and adjusted datasets, we estimated that unadjusted case  
 500 notifications are capturing on average 26.5% of symptomatic malaria cases in the district. This

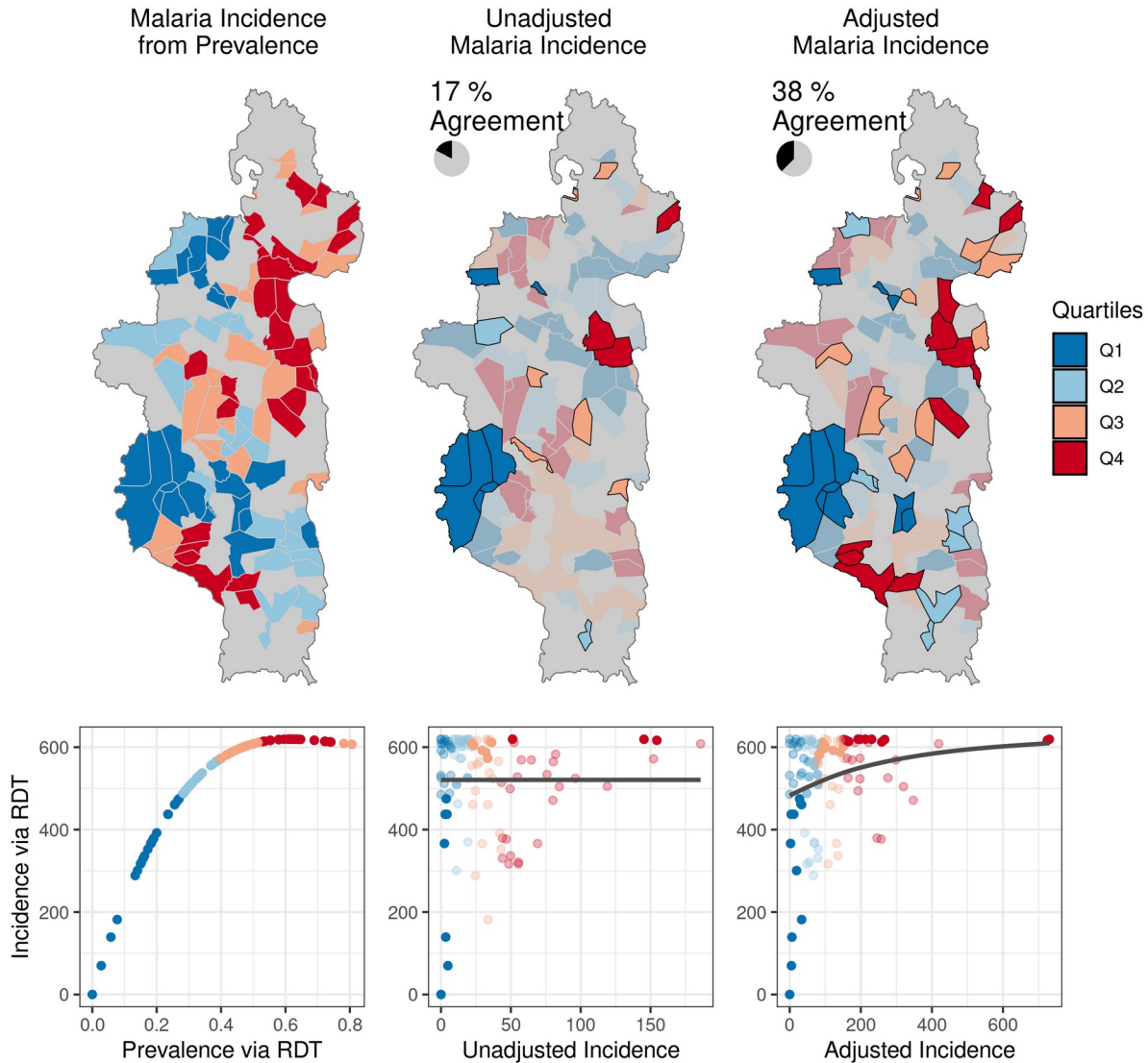
501 differed by year, with the lowest percentage of 24.2% in 2016 and the highest of 31.6% in 2017.  
502 The level of under-ascertainment also varied across fokontany. On average, the adjusted  
503 annual incidence in a fokontany was 9.15 (range: 1- 451) times the unadjusted annual  
504 incidence rate. However, when this was calculated omitting fokontany and year combinations  
505 that reported zero malaria cases in a year (26 out of 944), this ratio was reduced to 8.42  
506 (range: 1 – 76.5).

507

#### 508 ***Validation with Prevalence Data***

509 We validated ZERO-G by comparing ZERO-G estimated incidence rates with incidence rates  
510 derived from the IHOPE prevalence survey in children under 15 years old (Fig. 4). Unadjusted  
511 incidence rates were negatively correlated with IHOPE incidence rates based on prevalence,  
512 but this correlation was not significant (Spearman's  $\rho = -0.141$ , p-value = 0.2). The unadjusted  
513 incidence rates had no correlation with the calculated incidence of symptomatic individuals in  
514 the IHOPE survey (Spearman's  $\rho = -0.050$ , p-value = 0.6). After adjusting the data, we found a  
515 positive correlation between ZERO-G and IHOPE incidence rates (Spearman's  $\rho = 0.316$ , p-  
516 value = 0.001). While the estimated correlation coefficient between incidence rates and the  
517 proportion of symptomatic and RDT positive children was positive in the adjusted dataset, it  
518 remained insignificant (Spearman's  $\rho = 0.188$ , p-value = 0.06). The adjusted dataset also  
519 doubled the number of correctly-ranked fokontany into quantiles that matched those from the  
520 prevalence data (Fig. 4). The adjusted dataset correctly ranked 43 of 104 fokontany, compared  
521 to 18 in the unadjusted dataset.

522



523

524 **Figure 4. The adjustment method results in monthly malaria incidence rates in 2021 that**  
 525 **more closely correspond to measures of malaria prevalence in children under 15 years**  
 526 **old.** Left: Malaria incidence derived from prevalence as measured by rapid-detection tests  
 527 (RDT) in children under 15 years old from the IHOPE cohort survey. Colors represent quartiles  
 528 from Q1 (lowest incidence) to Q4 (highest incidence). The scatter plot illustrates the non-linear  
 529 relationship between prevalence and incidence. Middle: Monthly malaria incidence in the  
 530 unadjusted dataset. Quartiles that match those in the prevalence data are highlighted in black.  
 531 The scatter plot illustrates the relationship between unadjusted incidence and IHOPE incidence.  
 532 Right: Monthly malaria incidence in the ZERO-G adjusted dataset. Quartiles that match those in  
 533 the prevalence data are highlighted in black. The scatter plot illustrates the relationship between  
 534 ZERO-G incidence and IHOPE incidence. Monthly incidence has been chosen to correspond to  
 535 the month in which the IHOPE survey was conducted for that fokontany.  
 536

537 **Simulated Endemic Disease Data**

538 The ZERO-G estimator reproduced simulated true incidence data when applied to simulated  
539 reported incidence data that contained reporting biases due to healthcare access. The ZERO-G  
540 adjusted dataset predicted the true monthly incidence rates with a RMSE of 54.5, compared to a  
541 RMSE of 94.67 when using the unadjusted dataset (Fig. S1.4). It was also more strongly  
542 correlated with the true incidence rates (Spearman's  $\rho=0.82$ ,  $p\text{-value}<0.001$ ), compared to the  
543 unadjusted dataset (Spearman's  $\rho=0.43$ ,  $p\text{-value}<0.001$ ) (Fig. S1.5). In addition, the ZERO-G  
544 adjusted dataset reduced biases due to geographic distance and fee reimbursement policies  
545 seen in the unadjusted dataset (Fig S1.6, Fig S1.7). The unadjusted incidence dataset exhibited  
546 a strongly negative correlation with increasing distance to the nearest health clinic (Spearman's  
547  $\rho= -0.666$ ,  $p\text{-value} <0.001$ ), which was reduced by over 30% in the ZERO-G adjusted dataset  
548 (Spearman's  $\rho= -0.465$ ,  $p\text{-value}<0.001$ ). The ratio of incidence in zones served by health clinics  
549 offering fee reimbursement to incidence in zones without this policy was only 1.11 in the ZERO-  
550 G adjusted dataset, compared to 1.93 in the reported dataset and 1.00 in the true dataset.

551

## 552 **DISCUSSION**

553 There is a critical need for routine surveillance systems to produce estimates at the spatial scale  
554 of individual communities so that control interventions can be targeted in collaboration with  
555 community health programs. However, HMIS data are rarely kept disaggregated at this scale  
556 and, when they are, they suffer from under-estimation of incidence that varies across space and  
557 time, preventing their usefulness for decision making . We developed an adjustment method  
558 that combines a gravity-model of healthcare access with an indirect estimator to create long-  
559 term routine surveillance data at the community-scale, adjusted for under-ascertainment due to  
560 uneven health care access. We demonstrated this method by applying it to field-collected  
561 malaria case notification data from 192 communities over 5 years of surveillance in a rural  
562 District of Madagascar. This method reduced geographical and financial bias in field-collected  
563 malaria incidence rates by 91% and 96%, respectively. In addition, we validated this method  
564 with two external, population-representative datasets and found strong agreement with self-  
565 reported healthcare access and malaria prevalence rates. We further assessed the  
566 generalizability of the ZERO-G estimator on a simulated dataset and found it nearly doubled the  
567 ability to reproduce true incidence rates. The ZERO-G estimator can obtain estimates that  
568 approximate long-term active surveillance data of common, endemic diseases at fine-spatial  
569 scales using only passive surveillance data.

570 ZERO-G greatly reduced bias in malaria incidence rates from a passive surveillance  
571 dataset in our case study. In Ifanadiana district, per capita health system utilization rates are

572 twice as high for fokontany within 5km of a health center than those further away <sup>11</sup>, and we  
573 found similar trends in the unadjusted malaria data (Fig. 3). Geographic bias in the malaria data  
574 was therefore primarily reduced by accounting for low sampling intensity at those fokontany  
575 further than 5 km from a PHC (Fig. 2). Financial costs represent a significant barrier to  
576 healthcare seeking, particularly for low-income communities, and differential user fee policies  
577 over time (e.g. implementation of universal health coverage) can result in healthcare access  
578 patterns changing as a function of this <sup>35,36</sup>. In Ifanadiana, the removal of user fees to patients  
579 (via reimbursement policies to PHCs) in part of the district led to a sudden and sustained 65%  
580 increase in utilization rates <sup>27</sup>. ZERO-G removed this bias, resulting in similar incidence rates  
581 regardless of when and where reimbursement policies were in place. ZERO-G also resulted in  
582 data that more accurately identified malaria prevalence hotspots and coldspots than the  
583 unadjusted data, performing twice as well. However, the adjusted dataset only correctly  
584 categorized 38% of fokontany into ranked quantiles, illustrating the difficulty in matching  
585 incidence data to prevalence data. While we accounted for the non-linear relationship between  
586 malaria incidence and prevalence in our evaluation of ZERO-G, we did not account for age-  
587 specific differences in symptomatic rates between children and juveniles <sup>37</sup>, which may have  
588 further skewed this comparison. Further, we only had access to one study of malaria prevalence  
589 at a spatial-scale finer than 5 x 5 km. Therefore, we were only able to assess our method's  
590 ability to reproduce spatial patterns in malaria burden, and not temporal patterns. However, our  
591 model results agree with national-level trends in malaria, which witnessed over a 40% increase  
592 in confirmed malaria cases in 2020 <sup>38</sup>, suggesting we are capturing temporal trends as well.

593 Unlike other methods, which rely on external datasets describing sampling intensity that  
594 are collected at coarse spatial resolutions and infrequently (e.g. DHS, MICS, or other survey  
595 data), ZERO-G uses data that match the spatial and temporal resolution of the case notification  
596 data. This allows it to retain the original spatial and temporal scales at which the data was  
597 collected while relying solely on public health and demographic data that is easily accessible to  
598 public health actors. Population data can be sourced at fine-scale administrative levels via  
599 national census data or via open-source datasets such as PopGrid <sup>39</sup>. As with all estimates of  
600 population-level indicators, the lack of high-quality population estimates (the “denominator  
601 problem”<sup>40</sup>) is an obstacle to estimating incidence rates and may lead to biased estimates.  
602 Information on PHC locations and services are collected by Ministries of Health or available via  
603 regional, open-source datasets (e.g. <sup>41</sup>). These data may not always be available on a monthly  
604 basis, particularly staffing data. In these cases, annual or static data may be substituted for  
605 monthly data, as demonstrated in the Madagascar case study. In the context of health



606 interventions, however, the ability to track monthly changes to policies or health infrastructure  
607 due to an external intervention is a benefit of the ZERO-G estimator over existing methods. We  
608 used a field-verified transport network created via OpenStreetMap to estimate the distance  
609 between a population and a PHC, which accurately represents patients' distance to PHCs <sup>25</sup>;  
610 however, these transportation networks are not globally available. When transportation  
611 networks are not available, open-source databases of populations' distances to PHCs and other  
612 services could serve as suitable substitutes (e.g. <sup>42,43</sup>). Finally, consultation rates are commonly  
613 tracked by health systems and are increasingly recorded via electronic health management  
614 information systems <sup>44,45</sup>, facilitating their use in these estimates.

615         ZERO-G differs from existing adjustment methods in several ways. First, it uses monthly  
616 estimates of sampling intensity in the indirect estimate step rather than data from annual or  
617 inter-annual population surveys. Most adjustment methods do not account for changes in  
618 healthcare seeking behavior due to seasonality or temporal shifts to the health system (e.g.  
619 climate-driven changes in access, changes in PHC staffing rates, clinic-level interventions), and  
620 are therefore limited to inference at an annual frequency <sup>46</sup>. This functionality of the ZERO-G  
621 method is particularly beneficial in the context of partial health system interventions, such as the  
622 adoption of new policies or technologies. Second, the resulting dataset is available at the same  
623 spatial scale at which it is collected, rather than spatially interpolated between points or  
624 aggregated to coarser resolutions. We build on work by Hyde et al. <sup>16</sup>, which proposed a similar  
625 indirect estimation adjustment method for malaria data that featured a monthly frequency at the  
626 scale of the community, but dealt with extreme low incidence values by spatially smoothing  
627 estimates between neighboring communities, introducing spatial structure into the adjusted  
628 dataset and removing existing natural variation. Because ZERO-G estimates are available at  
629 the community level at a monthly frequency, they can be used to inform community health  
630 programs and spatially targeted interventions at the village level in real-time, capabilities that  
631 are lacking in other adjustment methods. In addition, ZERO-G explicitly models the sampling  
632 intensity as a function of geographic and health-system characteristics in all the facilities  
633 surrounding a community via a gravity model instead of using information from the closest  
634 facility in a linear model, as in Hyde et al. <sup>16</sup>. Because of this, changes in the health system,  
635 such as the closing of a facility due to a natural disaster or a policy change, can be directly  
636 incorporated into calculations of sampling intensity in near real-time. It also allows for estimation  
637 of sampling intensity in unsampled communities or months through these modeled processes,  
638 rather than relying on interpolation.

639           There are several limitations that should be taken into consideration when implementing  
640 ZERO-G. First, the adjustment of zero-incidence samples due to extremely low ascertainment  
641 introduces a further source of uncertainty. However, the identification of which samples to  
642 impute is data-driven, and, as demonstrated when applied to both the simulated and field-  
643 derived datasets, replaces only a small fraction of the overall data. Secondly, the ZERO-G  
644 estimator does not include a step to disaggregate consultation rates to a finer spatial scale than  
645 that reported by the PHC, often a major limiting step to accessing disease incidence data at a  
646 fine spatial scale. In Ifanadiana, the standard reporting system aggregates consultations at the  
647 level of the health facility catchment. We manually digitized health registers to obtain  
648 community-level data, a time- and resource-intensive process. However, the increased  
649 availability of electronic systems at the level of primary and community health care represents  
650 an opportunity to apply this method directly and in real time to data at fine spatial scales. Finally,  
651 the ZERO-G method is not appropriate for all passive case notification datasets. It is best suited  
652 for routine passive surveillance of common, endemic diseases, which possess the historical  
653 datasets needed to impute low-incidence values. The ZERO-G method is also inappropriate for  
654 adjusting case notifications of novel diseases because behavioral and health-system responses  
655 to a rapidly-evolving epidemic will violate the assumption that the relationship between  
656 healthcare access and sampling intensity of the disease is constant.

657           In conclusion, ZERO-G represents a promising new method for adjusting passive  
658 surveillance data of endemic diseases for under-ascertainment bias in regions with low and  
659 heterogeneous healthcare seeking rates, developed specifically for use at the community level.  
660 Unlike other methods, it is applicable in regions with ongoing heterogeneous public health  
661 interventions, allowing it to be used to adjust case notifications used in monitoring and  
662 evaluation efforts in addition to routine monitoring of diseases. This method can serve as part  
663 of a wider toolkit of statistical techniques used to improve targeted health system responses. In  
664 a case study in a rural health district in Madagascar, it was able to reduce geographic and  
665 financial bias in malaria incidence and the resulting dataset more closely approximated spatial  
666 trends in malaria prevalence. It is particularly suited to rural areas, where geographic isolation  
667 strongly influences healthcare access <sup>42</sup>. As spatially-explicit health metrics become an  
668 increasingly important tool for precision public health interventions, there is an urgent need to  
669 obtain and use quality data sources at the community scale. Statistical methods such as ZERO-  
670 G can be an important tool to support the role of community health programs in the local  
671 targeting of interventions for disease control.

672

673 **DATA AVAILABILITY STATEMENT**

674 All code and data needed to reproduce this study are available in a figshare repository (doi:  
675 10.6084/m9.figshare.22154492).

676

677 **ACKNOWLEDGMENTS**

678 We would like to thank the health professionals who collect passive surveillance data in addition  
679 to serving their patients and the Pivot data collection team for their work collecting and digitizing  
680 this data. We would also like to thank Ann Miller and Marius Randriamanambintsoa for their  
681 support of the IHOPE longitudinal survey.

682

683 REFERENCES

1. Murray, C. J. L., Lopez, A. D. & Wibulpolprasert, S. Monitoring global health: time for new solutions. *BMJ* **329**, 1096–1100 (2004).
2. Tatem, A. J. *et al.* Mapping populations at risk: improving spatial demographic data for infectious disease modeling and metric derivation. *Popul Health Metrics* **10**, 8 (2012).
3. Amouzou, A., Faye, C., Wyss, K. & Boerma, T. Strengthening routine health information systems for analysis and data use: a tipping point. *BMC Health Services Research* **21**, 618 (2021).
4. Rifkin, S. B. Alma Ata after 40 years: Primary Health Care and Health for All—from consensus to complexity. *BMJ Global Health* **3**, e001188 (2018).
5. World Health Organization, PEPFAR & UNAIDS. *Task shifting : rational redistribution of tasks among health workforce teams : global recommendations and guidelines.* (2007).
6. *Health for the people: National community health worker programs from Afghanistan to Zimbabwe.* (USAID MCHIP, 2020).
7. Young, M., Wolfheim, C., Marsh, D. R. & Hammamy, D. World Health Organization/United Nations Children's Fund Joint Statement on Integrated Community Case Management: An Equity-Focused Strategy to Improve Access to Essential Treatment Services for Children. *The American Journal of Tropical Medicine and Hygiene* **87**, 6–10 (2012).
8. Kok, M. C. *et al.* How does context influence performance of community health workers in low- and middle-income countries? Evidence from the literature. *Health Research Policy and Systems* **13**, 13 (2015).
9. Hodgins, S. *et al.* Community health workers at the dawn of a new era: 1. Introduction: tensions confronting large-scale CHW programmes. *Health Research Policy and Systems* **19**, 109 (2021).
10. Admon, A. J. *et al.* Assessing and improving data quality from community health workers: a successful intervention in Neno, Malawi. *Public Health Action* **3**, 56–59 (2013).
11. Garchitorena, A. *et al.* Geographic barriers to achieving universal health coverage: evidence

- from rural Madagascar. *Health Policy and Planning* **36**, 1659–1670 (2021).
12. Afrane, Y. A., Zhou, G., Githeko, A. K. & Yan, G. Utility of Health Facility-based Malaria Data for Malaria Surveillance. *PLOS ONE* **8**, e54305 (2013).
  13. Ohrt, C. *et al.* Information Systems to Support Surveillance for Malaria Elimination. *Am J Trop Med Hyg* **93**, 145–152 (2015).
  14. Noor, A. M., Zurovac, D., Hay, S. I., Ochola, S. A. & Snow, R. W. Defining equity in physical access to clinical services using geographical information systems as part of malaria planning and monitoring in Kenya. *Trop Med Int Health* **8**, 917–926 (2003).
  15. Cibulskis, R. E., Aregawi, M., Williams, R., Otten, M. & Dye, C. Worldwide Incidence of Malaria in 2009: Estimates, Time Trends, and a Critique of Methods. *PLOS Medicine* **8**, e1001142 (2011).
  16. Hyde, E. *et al.* Estimating the local spatio-temporal distribution of malaria from routine health information systems in areas of low health care access and reporting. *International Journal of Health Geographics* **20**, 8 (2021).
  17. Delamater, P. L., Shortridge, A. M. & Kilcoyne, R. C. Using floating catchment area (FCA) metrics to predict health care utilization patterns. *BMC Health Serv Res* **19**, 144 (2019).
  18. Hickman, M. & Taylor, C. Indirect Methods to Estimate Prevalence. in *Epidemiology of Drug Abuse* (ed. Sloboda, Z.) 113–131 (Springer-Verlag, 2005). doi:10.1007/0-387-24416-6\_8.
  19. Luo, W. & Qi, Y. An enhanced two-step floating catchment area (E2SFCA) method for measuring spatial accessibility to primary care physicians. *Health & Place* **15**, 1100–1107 (2009).
  20. Delamater, P. L. Spatial accessibility in suboptimally configured health care systems: A modified two-step floating catchment area (M2SFCA) metric. *Health & Place* **24**, 30–43 (2013).
  21. Mersmann, O. mco: Multiple Criteria Optimization Algorithms and Related Functions. (2020).
  22. Audigier, V. & Resche-Rigon, M. micemd: Multiple Imputation by Chained Equatoins with

- Multilevel Data. (2023).
23. Khan, A. A. An integrated approach to measuring potential spatial access to health care services. *Socio-Economic Planning Sciences* **26**, 275–287 (1992).
  24. Seidu, A.-A. Mixed effects analysis of factors associated with barriers to accessing healthcare among women in sub-Saharan Africa: Insights from demographic and health surveys. *PLOS ONE* **15**, e0241409 (2020).
  25. Ihantamalala, F. A. *et al.* Improving geographical accessibility modeling for operational use by local health actors. *International Journal of Health Geographics* **19**, 27 (2020).
  26. World Health Organization. *Global Health Expenditure Database*. <https://apps.who.int/nha/database> (2022).
  27. Garchitorea, A. *et al.* In Madagascar, Use Of Health Care Services Increased When Fees Were Removed: Lessons For Universal Health Coverage. *Health Affairs* **36**, 1443–1451 (2017).
  28. Cordier, L. F. *et al.* Networks of Care in Rural Madagascar for Achieving Universal Health Coverage in Ifanadiana District. *Health Systems & Reform* **6**, e1841437 (2020).
  29. Miller, A. C. *et al.* Cohort Profile: Ifanadiana Health Outcomes and Prosperity longitudinal Evaluation (IHOPE). *International Journal of Epidemiology* **47**, 1394–1395e (2018).
  30. R Core Team. R: A language and environment for statistical computing. (2021).
  31. Buuren, S. van & Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* **45**, 1–67 (2011).
  32. Clifford, P., Richardson, S. & Hemon, D. Assessing the Significance of the Correlation between Two Spatial Processes. *Biometrics* **45**, 123–134 (1989).
  33. Cameron, E. *et al.* Defining the relationship between infection prevalence and clinical incidence of Plasmodium falciparum malaria. *Nat Commun* **6**, 8170 (2015).
  34. Dalrymple, U. *et al.* The contribution of non-malarial febrile illness co-infections to Plasmodium falciparum case counts in health facilities in sub-Saharan Africa. *Malar J* **18**, 195 (2019).

35. Yates, R. Universal health care and the removal of user fees. *The Lancet* **373**, 2078–2081 (2009).
36. James, C., Morris, S. S., Keith, R. & Taylor, A. Impact on child mortality of removing user fees: simulation model. *BMJ* **331**, 747–749 (2005).
37. Lindblade, K. A., Steinhardt, L., Samuels, A., Kachur, S. P. & Slutsker, L. The silent threat: asymptomatic parasitemia and malaria transmission. *Expert Review of Anti-infective Therapy* **11**, 623–639 (2013).
38. World Health Organization. *World malaria report 2021*. (World Health Organization, 2021).
39. Leyk, S. *et al.* The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use. *Earth System Science Data* **11**, 1385–1409 (2019).
40. Morrison, C. N. *et al.* The unknown denominator problem in population studies of disease frequency. *Spat Spatiotemporal Epidemiol* **35**, 100361 (2020).
41. Maina, J. *et al.* A spatial database of health facilities managed by the public health sector in sub Saharan Africa. *Sci Data* **6**, 1–8 (2019).
42. Weiss, D. J. *et al.* Global maps of travel time to healthcare facilities. *Nat Med* **26**, 1835–1838 (2020).
43. Nelson, A. *et al.* A suite of global accessibility indicators. *Sci Data* **6**, 266 (2019).
44. Kumar, M., Gotz, D., Nutley, T. & Smith, J. B. Research gaps in routine health information system design barriers to data quality and use in low- and middle-income countries: A literature review. *The International Journal of Health Planning and Management* **33**, e1–e9 (2018).
45. Siyam, A. *et al.* The burden of recording and reporting health data in primary health care facilities in five low- and lower-middle income countries. *BMC Health Services Research* **21**, 691 (2021).
46. Gibbons, C. L. *et al.* Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. *BMC Public Health* **14**, 147 (2014).

47. Gething, P. W. *et al.* Improving Imperfect Data from Health Management Information Systems in Africa Using Space–Time Geostatistics. *PLOS Medicine* **3**, e271 (2006).