

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

4
5 Michelle V Evans^{1,2,3*}, Felana A Ihantamalala^{2,3}, Mauricianot Randriamihaja^{1,2}, Andritiana
6 Tsirinomen'ny Aina², Matthew H Bonds^{2,3}, Karen E Finnegan^{2,3}, Rado JL Rakotonanahary^{2,3},
7 Mbolatiana Raza-Fanomezanjanahary², Bénédicte Razafinjato², Oméga Raobela⁴,
8 Sahondraritera Herimamy Raholiarimanana⁴, Tiana Harimisa Randrianaivalona⁴, Andres
9 Garchitorena^{1,2}

10
11
12
13
14
15
16
17

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

2. NGO Pivot, Ranomafana, Ifanadiana, Madagascar

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

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

18 **Table of Contents**

- 19 1. Creating Simulated Incidence Data
20 2. Case Study: Malaria Incidence in Ifanadiana, Madagascar
21 3. Reflexivity Statement

22
23

24 **1. Creating Simulated Incidence Data**

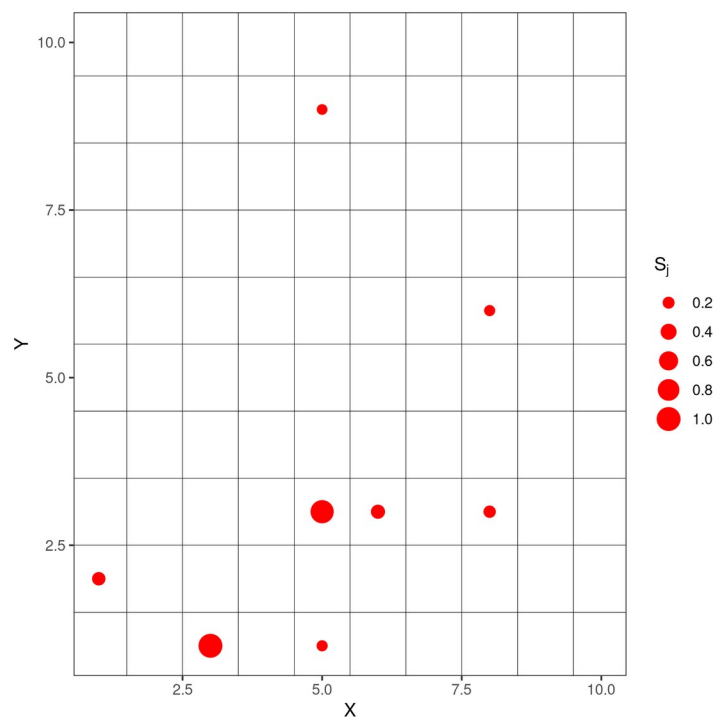
25 *Spatial distribution of health care infrastructure*

26 Monthly disease incidence was simulated for 100 administrative zones (patches, p) over 5
27 years. The patches were distributed in a 10 x 10 square matrix representing a health district
28 (Figure S1.1). Each patch's population was drawn from a uniform distribution between 800 and
29 1200 and age-stratification was not considered. The population remained constant over the
30 simulated time period.

31

32 Eight primary health clinics (j) were randomly distributed across the 10 x 10 matrix (Figure S1).
33 Clinics differed in the number of staff (random uniform from 5 to 15), whether they offered
34 advanced services (randomly distributed so that 50% of clinics offered advanced services), and
35 whether health care was provided free-of-charge at the clinic (randomly distributed so that 50%
36 of both advanced and basic clinics offered this service).

37



38

39 **Figure S1.1. Distribution of primary health clinics (red points) distributed among a**
40 **matrix of 10 x 10 administrative zones (squares).** The size of the point for each clinic
41 corresponds to the level of services it provides (S_j).
42

43

43 *Disease Dynamics*

44 We simulated consultation rates for constant background disease rates and for two diseases
45 that exhibited annual seasonality in their burdens. We assumed the background disease rate

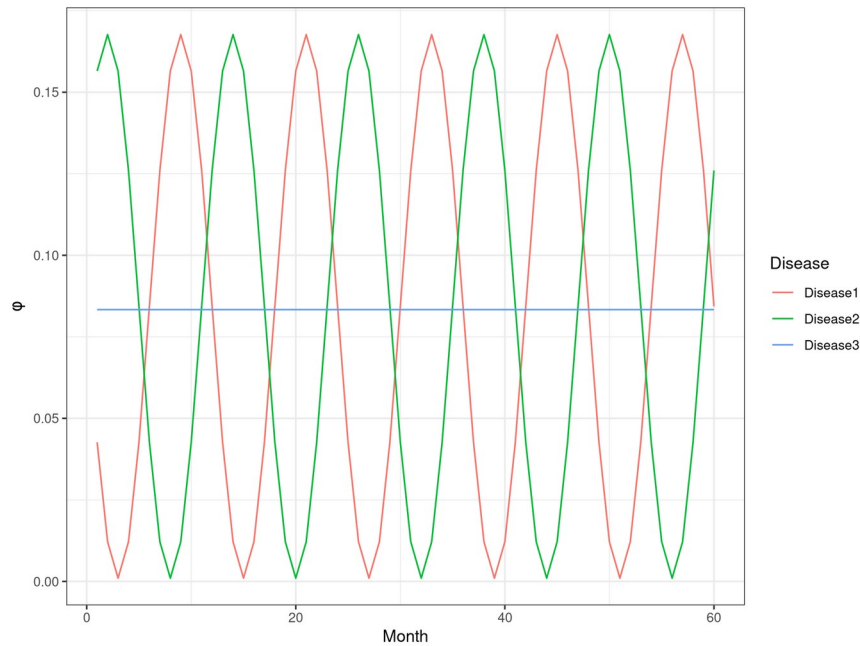
46 was one infection per person per year. We set the annual incidence of each seasonal disease to
 47 one infection per person per year, but varied the seasonality of each disease separately,
 48 resulting in a monthly risk of infection for each disease g (ϕ_g) (Figure S1.2). Each individual's
 49 probability of infection for each disease during each month was defined as the inverse logit of
 50 the logit-transformed monthly risk of infection plus a normally distributed random error with a
 51 mean of 0 and standard deviation of 0.5, resulting in a probability ranging from 0 - 1. This extra
 52 error was added so that the simulated data approximated the noisiness seen in field-derived
 53 disease notification data. This resulted in a number of cases for each disease g in patch p during
 54 month t drawn from a binomial distribution of size equal to the patch's population and
 55 probability ϕ_g (Eq. S1).

56

57

$$C_{g,p,t} = \text{Bin}(\text{population}_p, \phi_g) \quad (\text{Equation S1})$$

58



59

60 **Figure S1.2. The monthly risk of infection (phi) for each of three diseases across the**
 61 **simulated time period.**

62

63 *Reporting Rate*

64 We modeled an individual's probability of seeking health care at the patch level (PC_p) as a
 65 function of the distance to health clinics and the characteristics of those clinics (Eq. S2).

66

67

$$PC_p = \sum_j S_j (e^{-0.3d_{pj}^2}) \quad (\text{Equation S2})$$

68

69 Where S_j is the services provided by each clinic j and d_{pj} is the distance between patch p and
70 clinic j . The services provided by each clinic j were a function of the number of staff of that clinic
71 ($x1$), whether it offered advanced services ($x2$), and whether healthcare was provided free of
72 charge ($x3$), scaled to range from 0 – 1 (Eq. S3).

73

$$74 \quad S_j = \frac{x1_j * (x2 + 1) * (x3 + 1)}{\max_{k \in [j]} S_j} \quad (\text{Equation S3})$$

75

76 In addition, we simulated instances of zero reported infections per patch for each disease due to
77 1) a combination of low reporting rates and low disease risk and 2) randomness. These two
78 causes of zero reported infections were simulated independently from each other by a random
79 binomial event given a corresponding probability of a zero. The probability of a zero due to low
80 reporting rates (PC_p) and low disease risk (ϕ_g) for each disease g in patch p at month t was
81 calculated following Equation S4:

$$82 \quad Pz_{g,p,t} = 1 - PC_p^{1-\phi_g^{0.1}} \quad (\text{Equation S4})$$

83

84 The probability of a zero due to randomness (Pzr) was set at a constant value of 0.1.

85

86 The number of reported monthly cases for each disease per patch was therefore defined as:

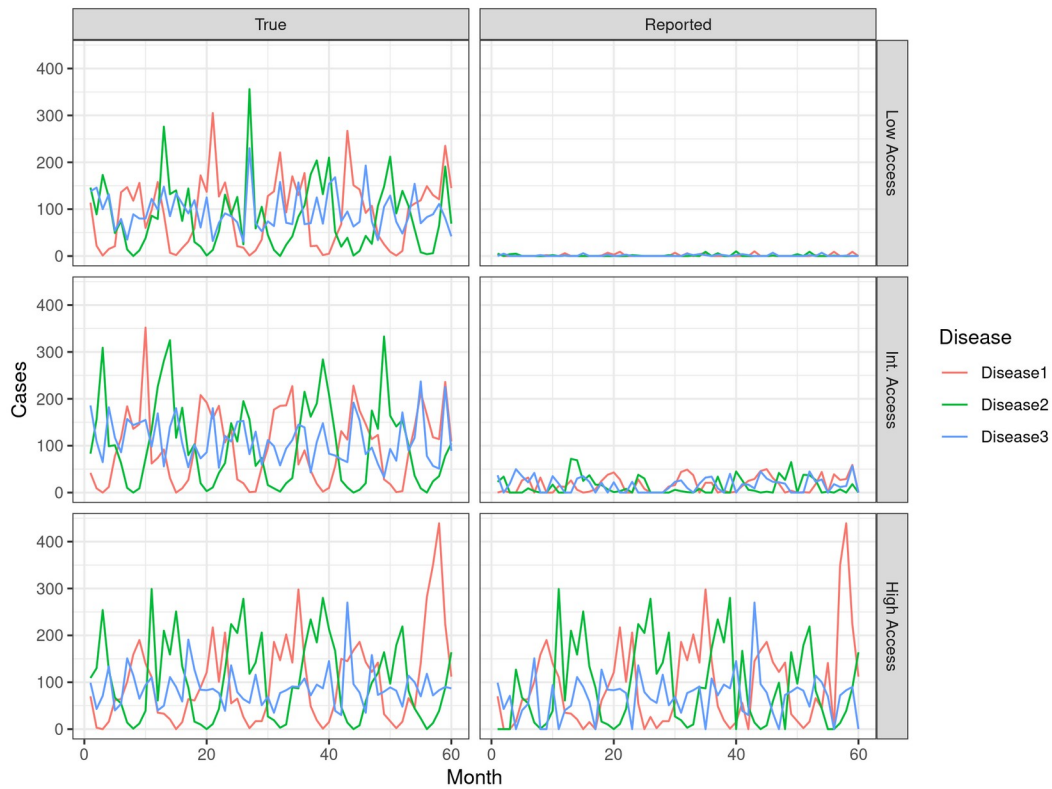
$$87 \quad R_{g,p,t} = Bin(C_{g,p,t}, PC_p) * Bin(1, (1 - Pz_{g,p,t}) * Bin(1, 1 - Pzr)) \quad (\text{Equation S5})$$

88

89

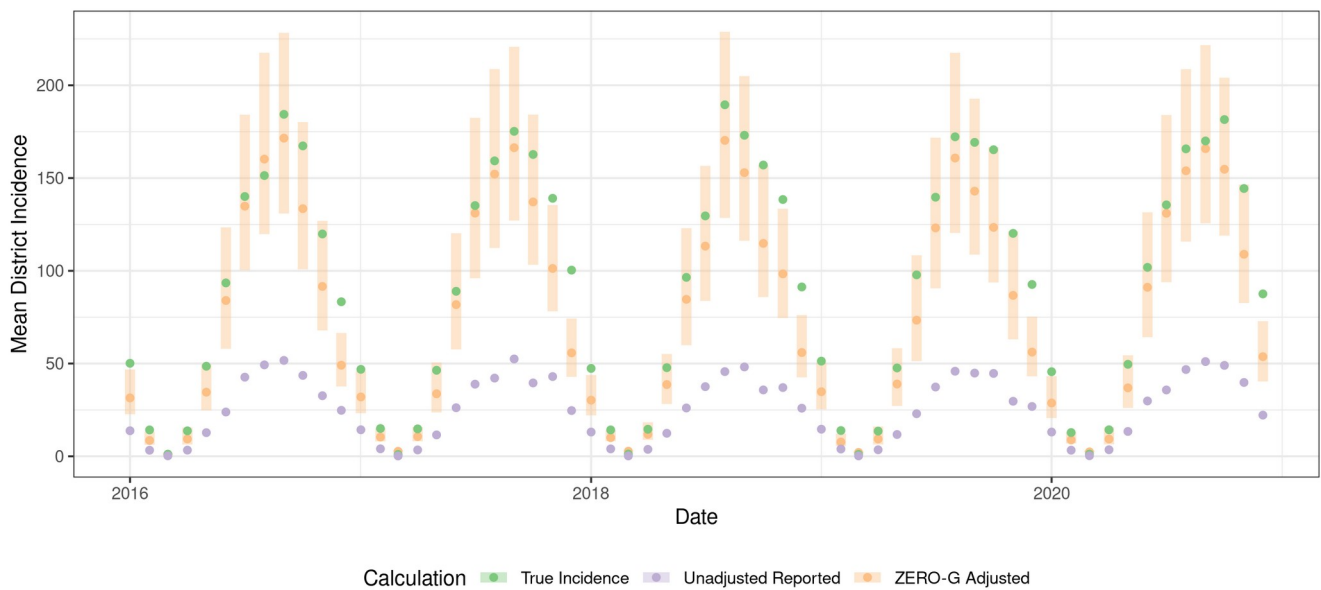
90 An example of comparing the true vs. reported cases for all three diseases for two patches of
91 differing health care access is shown in Figure S1.3. Notably, this dataset bears characteristics
92 that resemble realistic passive notification datasets, including a high variance around the mean
93 and unexplained missing data reported as zeros.

94

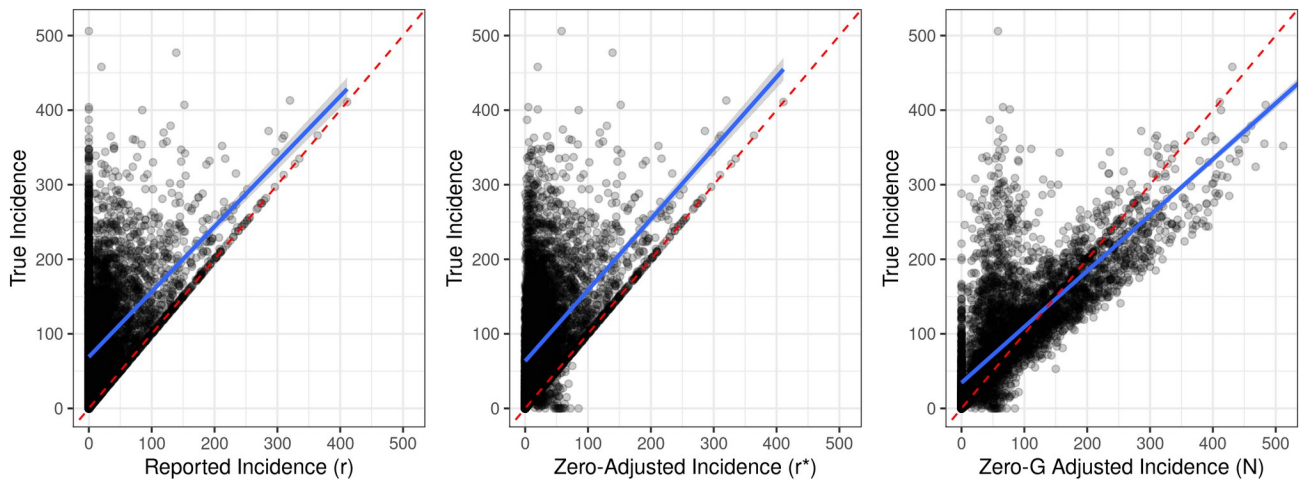


95
96
97
98

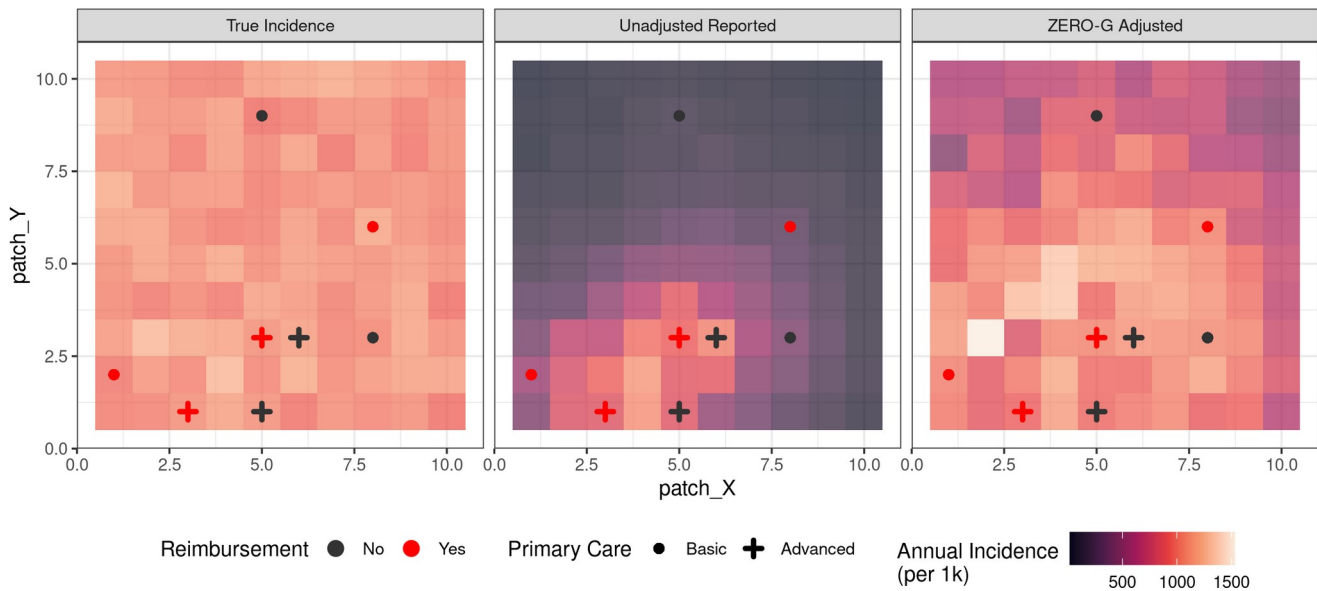
Figure S1.3. The true number of cases and reported number of cases for all three diseases in three example patches with low, intermediate, and high probability of seeking healthcare.



100 **Figure S1.4.** Time series of district-level incidence rates in the simulated dataset, for the true
101 dataset, reported dataset, and ZERO-G adjusted dataset. The bars represent the 95%
102 confidence intervals of the ZERO-G adjustment.

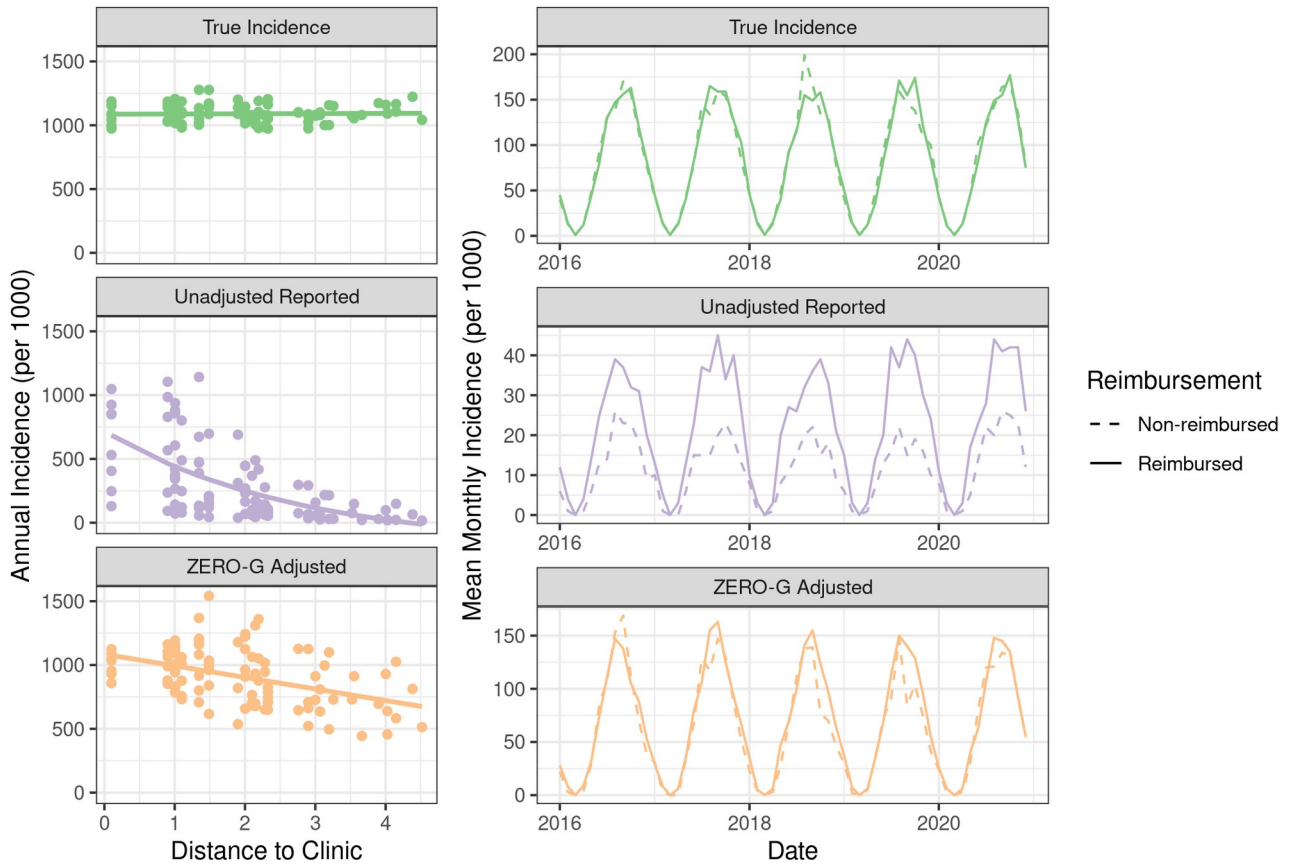


104 **Figure S1.5** Scatter plots comparing simulated true incidence data with unadjusted incidence
 105 rates, incidence rates adjusted for erroneous zeros, and ZERO-G adjusted incidence estimates.
 106



108 **Figure S1.6.** Spatial pattern of mean annual incidence per administrative zone in the true
 109 incidence dataset, reported dataset, and the ZERO-G adjusted dataset. The annual incidence is
 110 represented by the shaded color of each zone and the location of health clinics are represented
 111 by the points. Health clinics offering advanced primary care are represented by a cross and
 112 those offering basic care are represented by a circle, with the color of the point corresponding to
 113 whether fees are reimbursed at that clinic.

114
 115



117

118 **Figure S1.7.** Biases due to geographic location and financial policies were reduced in the
 119 ZERO-G adjusted data relative to the unadjusted data. Left: The average annual incidence per
 120 patch relative to a zone’s distance to the nearest clinic. Right: The mean monthly incidence in
 121 zones with fee reimbursement and zones without fee reimbursement policies.

122

123

124 **2. Case Study: Malaria Incidence in Ifanadiana, Madagascar**

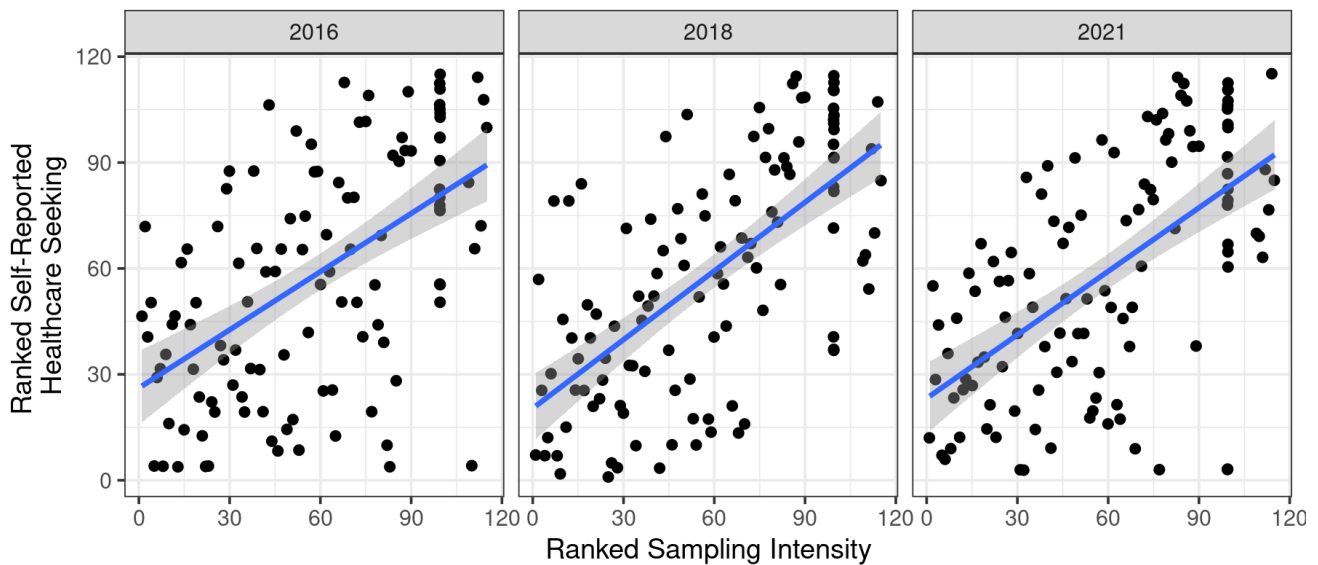
125 **Table S2.1.** Table of best fit parameters estimated via MLE in the estimation of healthcare
 126 access A via Equations 7- 12.

Parameter	Description	Age Class		
		Children	Juvenile	Adult
β_{s1}	PHC coefficient: initial Pivot intervention	9.360	9.124	25.248
β_{s2}	PHC coefficient: fee reimbursement	17.274	19.163	34.732
β_{s3}	PHC coefficient: PHC type	0.049	0.099	-0.285
β_{s4}	PHC coefficient: Number of staff	2.069	1.386	2.584
β_{s5}	PHC coefficient: Distance to District	6.475	2.833	3.743
λ	Distance decay	0.117	0.109	0.096
β_{λ}	Distance decay shape parameter	-0.278	-0.075	-0.26
β_C	Competition for PHC services coefficient	0.090	0.068	0.058
β_{g1}	Time coefficient: linear trend	-0.003	-0.003	0.002
β_{g2}	Time coefficient: seasonal trend	0.369	-0.572	0.168
ϕ	Time coefficient: month offset	12.801	7.477	-9.633

127

128

129



130

131 **Figure S2.1** Scatter plots illustrating the relationship between sampling intensity estimated via
 132 the floating catchment area (FCA) method and the self-reported healthcare seeking rates from
 133 the IHOPE cohort for three survey years (2016,2018,2021).