

1 **Using simulated infectious disease outbreaks to guide the design of individually randomized vaccine**
2 **trials**

3 Zachary J. Madewell^a, Ana Pastore Y Piontti^b, Qian Zhang^b, Nathan Burton^c, Yang Yang^a, Ira M.

4 Longini^a, M. Elizabeth Halloran^{d,e}, Alessandro Vespignani^b, Natalie E. Dean^a

5

6 ^a Department of Biostatistics, University of Florida, Gainesville, FL, United States

7 ^b Laboratory for the Modeling of Biological and Socio-technical Systems, Northeastern University,

8 Boston, MA, United States

9 ^c Institute for Child Health Policy, University of Florida College of Medicine, Gainesville, FL, United

10 States

11 ^d Biostatistics, Bioinformatics, and Epidemiology Program, Vaccine and Infectious Diseases Division,

12 Fred Hutchinson Cancer Research Center, Seattle, WA, United States

13 ^e Department of Biostatistics, University of Washington, Seattle, WA, United States

14

15 **Corresponding author**

16 Zachary J. Madewell, PhD, MPH

17 Department of Biostatistics

18 University of Florida

19 PO Box 117450

20 2004 Mowry Road

21 Gainesville, FL 32611-7450

22 Email: zmadewell@ufl.edu

23

24 **Short title:** Simulation data to design vaccine trials

25 **Word count:** 2808

26 **Funding:** This work was supported by the National Institutes of Health R01-AI139761.

27 **Abstract**

28 **Background/Aims:** Novel strategies are needed to make vaccine efficacy trials more robust given the
29 uncertain epidemiology of outbreaks. Spatially resolved mathematical and statistical models can help
30 investigators identify sites at highest risk of future transmission and prioritize these for inclusion in trials.
31 Models can also characterize the uncertainty in whether transmission will occur at a site, and how nearby
32 or connected sites may have correlated outcomes. A structure is needed for how trials can use models to
33 address key design questions, including how to prioritize sites, the optimal number of sites, and how to
34 allocate participants across sites.

35 **Methods:** We illustrate the added value of models using the motivating example of Zika vaccine trial
36 planning during the 2015-2017 Zika epidemic. We used a stochastic, spatially resolved, agent-based
37 transmission model (GLEAM) to generate 1,142 epidemics and site-level incidence at 100 high-risk sites
38 in the Americas. We considered several strategies for prioritizing sites (average site-level incidence of
39 infection across epidemics, median incidence, probability of exceeding 1% incidence), selecting the
40 number of sites, and allocating sample size across sites (equal enrollment, proportional to average
41 incidence, proportional to rank). To evaluate each design, we stochastically simulated trials in each
42 hypothetical epidemic by drawing observed cases from site-level incidence data.

43 **Results:** When constraining the overall trial sample size, the optimal number of sites represents a balance
44 between prioritizing highest-risk sites and having enough sites to reduce the chance of observing too few
45 endpoints. The optimal number of sites remained roughly constant despite varying the targeted number of
46 events, although it is necessary to increase the total sample size to achieve the desired power. Though
47 different ranking strategies returned different site orders, they performed similarly with respect to trial
48 power. Instead of enrolling participants equally from each site, investigators can allocate participants
49 proportional to projected incidence, though this did not provide an advantage in our example because the
50 top sites had a roughly similar risk profile. Sites from the same geographic region may have similar
51 outcomes, so optimal combinations of sites may be those that are more geographically dispersed, even
52 when these are not the highest ranked sites.

53 **Conclusions:** Mathematical and statistical models may assist in the design of successful vaccination trials
54 by capturing uncertainty and correlation in future transmission. Although many factors affect site
55 selection, such as logistical feasibility, models can help investigators optimize site selection and the
56 number and size of participating sites.

57

58 **Keywords:** clinical trial design; simulations; mathematical modeling; infectious diseases; vaccine;
59 forecast model; trial planning

60 **Background**

61 To observe enough events to reliably measure the efficacy of a vaccine, phase III trials often
62 enroll thousands or tens of thousands of participants across multiple sites. For endemic diseases like
63 rotavirus or malaria, incidence may be low but is relatively predictable. Investigators can use historical
64 data to guide the selection of trial populations assuming that future trends will be similar. Where
65 incidence is lower than expected during the trial, investigators can expand the sample size at existing sites
66 or increase participant follow-up to compensate. This strategy is unlikely to work for outbreak pathogens.
67 Historical data may be only weakly predictive of future incidence at a location. In fact, for pathogens with
68 high attack rates, an area with a large prior outbreak may be less susceptible to a subsequent outbreak if
69 there is a build-up of population immunity. Alternatively, that area may be more prone to another
70 outbreak if immunity wanes or the number of susceptible individuals is replenished. The outbreaks
71 themselves are highly unpredictable – when and where they will occur, how many will become infected,
72 and how long they will last. The 2014-2016 West African Ebola epidemic was emblematic of this
73 challenge, with a Phase III trial in Liberia enrolling over 8,000 individuals but observing no events
74 because the local outbreak subsided.¹ In this situation, expanding enrollment at existing sites or extending
75 follow-up of participants would not be able to compensate.

76 Novel strategies are needed to make vaccine trials more robust to the uncertain epidemiology of
77 outbreaks.² One recommended approach is to enable the addition of new sites over time using a master or
78 core protocol framework.³ If transmission in early hotspots is brought under control before the study has
79 reached a conclusion, the trial can continue at new hotspots. If the outbreak is declared over, the trial can
80 be paused until a subsequent outbreak. Spatially resolved mathematical and statistical forecast models can
81 assist investigators in selecting participating sites. Models can incorporate site-specific features such as
82 population size and density, socioeconomic vulnerability, sociocultural acceptance, logistic feasibility,
83 prior immunity estimated from traditional surveillance or serosurveys, ongoing local transmission, or risk
84 of importation. For vector-borne diseases, models can capture vector presence or abundance, sensitivity to
85 temperature and humidity, the spread of other diseases by the same vector, and whether other diseases

86 interfere with the disease of interest. By integrating diverse data sources, models can help investigators
87 identify sites at highest risk of future transmission and prioritize these for inclusion in the trial.⁴

88 Another advantage of simulation models for infectious disease trials is that they enable
89 investigators to explore a range of trial design features.⁵ Projected incidence is important, but so is the
90 uncertainty around that projection, including the probability of no or little future spread. When there is a
91 chance that sites will have little or no transmission, it becomes more important to include multiple,
92 geographically dispersed sites, to distribute this risk.

93 We illustrate the potential role of forecast modeling by using simulation data from a stochastic,
94 highly spatially resolved, agent-based Zika virus (ZIKV) model⁶ that was used to inform Zika vaccine
95 trial planning in 2016.⁷ Although the Zika epidemic subsided so that vaccine efficacy trials were not
96 possible,⁸ these are the type of data investigators would have at their disposal when designing future
97 efficacy trials for other infectious diseases. In addition to generalizable findings, we provide a plan for
98 how future trials may analyze their modeling results to prioritize test sites, site size, and the total number
99 of sites. We explore how disease models can be used to address key trial design questions, including how
100 to rank sites, the optimal number of sites to include, and how to allocate participants across sites.
101 Simulations can also be used to explore trial feasibility given financial, logistical, or time constraints. We
102 further consider how correlation between sites due to geographic proximity or human movement impacts
103 trial power.

104

105 **Simulation structure**

106 *Model*

107 We used the Global Epidemic and Mobility model (GLEAM) to identify the top 100 sites in the
108 Americas with the highest projected ZIKV probability of transmission and infection rates in 2017. These
109 projections were prepared in 2016, reflecting the type of data available to investigators planning trials.
110 GLEAM, which has been described elsewhere,^{6,9} is a discrete stochastic epidemic computational model
111 incorporating high-resolution demographic, socioeconomic, temperature, vector occurrence probability,

112 and human mobility data. The projections were calculated using discrete time steps of one day to simulate
113 transmission dynamics, but the results are summarized as number of infections per month. The resulting
114 dataset included the site name, population size, and number of simulated infections (both symptomatic
115 and asymptomatic) by month for 1,142 simulated epidemics from January through December 2017 (Table
116 S1). Population sizes for sites included all ages. Thus, we can examine both the range of projections for
117 each site, as well as look across sites within an epidemic.

118

119 *Trial design*

120 We describe the design of a hypothetical individually randomized Zika vaccine efficacy trial.
121 The primary outcome is total number of confirmed symptomatic ZIKV cases. Given a set of selected sites
122 and a fixed enrolled population for each site, for which we consider various different combinations, we
123 simulate a trial as follows. First, we select one of the 1,142 simulated epidemics, which has an associated
124 annual infection attack rate for each site. We simulate the number of infected trial participants at each site
125 as a binomial draw with the probability of infection set at the site-level attack rate, and then we draw the
126 number of these with symptomatic disease assuming 20% symptomatic proportion.¹⁰ This yields the total
127 number of cases at each site, which is then added across sites. We repeat the binomial draws 50 times at
128 each site, and then across all 1,142 epidemics to generate 57,100 simulated trials.

129 Approximately 60 symptomatic infection events are needed to have 90% power to reject the null
130 hypothesis that the vaccine efficacy $\leq 30\%$ when it is actually 70% using a 1:1 allocation to vaccine or
131 placebo. We therefore defined a successful (i.e., adequately powered) trial as finding ≥ 60 cases across all
132 sites in one year; we also explored trial designs targeting 50 to 150 events.

133 In a sensitivity analysis, we consider the feasibility of trials when attack rates are uniformly lower
134 than projected by the model. To explore this scenario, we restrict analyses to the 25% of simulated
135 epidemics with lowest overall infection attack rates across all sites.

136

137 **Findings**

138 *Number of sites*

139 The first key design choice is the number of participating sites. Sites are ranked by mean
140 incidence of infection across all simulated epidemics (Figure S1), and we consider designs including the
141 top site, the top two sites, and so on. For this example, we constrain the overall sample size at 15,000
142 participants and allocate these participants equally across selected sites. We plot the distribution of the
143 simulated number of cases for each design in Figure 1. Starting on the left side of the figure, the bimodal
144 nature of outbreaks is apparent when five or less sites are included. While the median number of cases of
145 the one-site design is highest relative to other designs, with a high upper tail observed for large outbreaks,
146 there is notable mass near zero cases, when little transmission occurs at the site. As the number of sites
147 increases, this bimodal phenomenon disappears; the probability of having zero cases decreases, but the
148 median expected number of cases also decreases because lower incidence sites are included.

149 While it is theoretically possible to enroll from only a single site, this presents an unacceptable
150 risk of failing to accrue the needed endpoints. It may also not be practically feasible if the site has a small
151 population. Furthermore, while a very high attack rate in a trial could shorten the trial duration or increase
152 study precision, our primary goal is to meet our target number of events, not dramatically exceed it. Thus,
153 rather than median expected number of cases, it is preferable to examine the probability that the design is
154 adequately powered. The curves in Figure 2 A plot the probability of success (here defined as exceeding
155 the target of 60 cases) as a function of the number of sites. We observe a local maximum around 8-11
156 sites, such that too few or too many sites are suboptimal with respect to trial success. In practice, the exact
157 location of this maximum will depend on the specific epidemiological setting.

158 The location of this local maximum for number of sites is reasonably stable to the ranking
159 criterion, even if we change the target number of events or total sample size. Figure 3 shows the minimum
160 trial size required to achieve at least 90% probability of success for different target numbers of events and
161 sites. As the target number of events increases, there is an expected increase in the total number of
162 participants needed. Yet for any specified target number of events, the desired 90% probability of success
163 is achieved with the smallest overall sample size when around eight sites are included.

164 The optimal design depends on the underlying simulation data through the site-level attack rates.
165 We consider a sensitivity analysis where the overall epidemic is smaller than projected by restricting to
166 the 25% of simulations with lowest overall infection attack rates across all sites. Figure 4 compares the
167 probability of success of different designs for all simulations versus the low incidence subset. While many
168 of the same relationships persists, the probability of success drops dramatically. Even increasing the total
169 number of sites and overall sample size, the probability of success does not exceed 25% in the designs
170 explored. Thus, this approach is also useful for exploring the feasibility of trials.

171 Figure 4 also demonstrates that the two-site design is best when the targeted trial size is relatively
172 small. Even for large targeted trial sizes, the difference in success probability between the two-site design
173 and more sites designs is no more than 10%. Logistically, the two-site design is an appealing option.

174

175 *Site prioritization*

176 In the previous section, sites are ranked by average model-projected site-level incidence of
177 infection. We examined other ranking strategies, including median model projected site-level incidence,
178 and the proportion of simulated outbreaks where site-level incidence exceeds a threshold, such as 1%.
179 The latter strategy is intended to capture the bimodal nature of outbreaks, and that a few very large
180 outbreaks could drive a high average incidence. In general, these measures are well-correlated, but they
181 can yield different rankings (Figure S2, Table S2). Small sites may have higher attack rates, but may also
182 have a higher probability of having zero infections across all participants. Nonetheless, we found similar
183 performance across the different ranking strategies (Figure 2).

184

185 *Allocation strategies*

186 Next, we considered different strategies for allocating the total sample size across multiple sites.
187 The strategies are to distribute enrollment size 1) evenly across all sites, 2) proportional to mean
188 incidence of infection, and 3) a middle-ground strategy using the average of sample sizes obtained from
189 the previous two strategies.

190 1) Equal enrollment: Let N be the total sample size, let m be the number of sites, let n_i be the
191 sample size at site i , then: $n_i = \frac{N}{m}$

192 2) Proportional to mean incidence: Let r_i be the rank at site i and let y_i = mean incidence. Then site
193 i has size: $n_i = N \frac{y_i}{\sum_{j:r_j \leq m} y_j}$

194 3) Average of equal enrollment and proportional to mean incidence: $n_i = \frac{\frac{N}{m} + \left(N \frac{y_i}{\sum_{j:r_j \leq m} y_j} \right)}{2}$

195
196 An example with five sites is shown in Table S3. The difference between the strategies will depend upon
197 how similar projected incidence is across the top-ranked sites.

198 In this example, enrolling participants proportional to average incidence did not outperform the
199 other strategies when fewer than 15 sites were included (Figure 2 B). This covers the range where the
200 probability of success maximizes, around 8-11 sites. With larger numbers of enrollment sites, enrolling
201 participants proportional to average incidence outperformed the other strategies as fewer individuals are
202 enrolled from sites that are expected to have lower attack rates. However, designs with more sites are sub-
203 optimal based on their reduced probability of success. As expected, the middle-ground strategy performs
204 in between the others, but it may be desirable for logistical reasons by balancing enrollment across sites.
205 The similar performance across the strategies for fewer than 15 sites may reflect that sites had similar
206 enough risk and large enough uncertainty that proportional allocation provided no worthwhile advantage.

207
208 *Correlation between sites*

209 It is conceivable that incidence rates are similar among sites in the same geographic region. Aside
210 from Brazil, Figure 5 shows that correlation in incidence is highest among sites from the same country. It
211 may therefore be that the sites with the highest simulated incidence are all from the same geographic
212 region. To reduce the chance of enrolling sites from only one geographic area that may, by chance, have a
213 smaller than expected outbreak, it may be prudent to simultaneously enroll participants from other
214 geographically dispersed sites.

215 We explored whether alternative combinations of sites (that may be more geographically
216 dispersed) could have a higher probability of success than those based only on rankings. Figure 6 displays
217 combinations of three, four, five, six, and seven site trials that achieved higher probability of success than
218 a design that selects sites based solely on average site-level incidence. For example, the top three sites
219 based on incidence are relatively close together, as seen by the low mean pairwise distance between sites
220 plotted on the X-axis. Many other combinations of three sites return higher probability of success, and
221 these tend to be more geographically dispersed (higher mean pairwise distance). A similar pattern is
222 observed for higher numbers of sites, although the gains become more modest.

223

224 *Number of countries*

225 While it may be prudent to recruit sites from at least a few different countries (under the intuition
226 that sites within a country are correlated), an important operational consideration is the number of
227 countries enrolled. For each country included, the logistical burden for the trial increases substantially
228 because it involves engaging with multiple ministries of health and the protocol needs to be approved by
229 country-level institutional review boards. Thus, investigators may prefer to pursue trials in countries that
230 have many high-risk sites. Figure S3 visualizes the number and incidence of sites by country. For
231 example, in this scenario, it may be practical to select several sites from Peru, Mexico, Ecuador, and
232 Colombia since many candidate sites are high risk.

233

234 **Conclusions**

235 We describe the use of simulation data for design elements for individually randomized vaccine
236 trials during an ongoing epidemic. Mathematical models allow us to capture a range of possible
237 outcomes, from small to large outbreaks, and incorporate correlation between sites connected by human
238 movement. Models therefore may capture the stochasticity of future transmission, reflecting a distribution
239 rather than only the mean projection, which may be used to guide trial planning.

240 We used a single model to identify the top 100 sites in the Americas with the highest projected
241 Zika virus transmission probability and infection rates in 2017 and leveraged those data to analyze
242 vaccine efficacy trial design strategies. However, in a real-world infectious disease outbreak, ensemble
243 forecast modeling (combining projections from independent modeling groups) may be used to
244 expeditiously guide identification of appropriate sites for vaccine efficacy trials.⁴ In practice,
245 implementation will depend upon practical considerations with regard to site selection, speed of rollout,
246 number of participants across sites, including a site's capacity to support a vaccine trial. There may also
247 be political considerations.

248 The optimal number of sites to enroll represents a balance between having enough sites to
249 distribute the risk of having a smaller than expected outbreak at any one site, versus enrolling participants
250 from lower risk sites. This optimal number of sites stayed relatively constant even when increasing the
251 targeted number of events, although the sample size must increase to achieve the desired probability of
252 success. Though different methods of prioritizing sites (average incidence, median incidence, and
253 probability of exceeding a threshold) returned different rankings, overall the approaches performed
254 similarly in this example. In general, they return the same set of prioritized sites. Investigators can use
255 simulations to guide allocation of participants across sites, prioritizing high-risk sites, though this appears
256 to provide an advantage over a simpler equal allocation approach only when there are large differences in
257 risk across sites.

258 We demonstrated that sites from a geographic region may have similar outcomes and that optimal
259 combinations of sites may be those that are geographically dispersed. Perhaps it is feasible to develop
260 novel search algorithms in the enormous space of site combinations without simulating trials for all
261 possible combinations. However, including multiple countries increases trial costs and logistical
262 complexities. Cost-effectiveness analyses may be used to explore the potential benefits and feasibility of
263 specific trial design features by assessing the net costs or savings of vaccinating specific populations.¹¹
264 Although many factors affect site selection, models can help investigators optimize site selection and the
265 number and size of participating sites.

266

267 **Declaration of conflicting interests**

268 APyP, QZ and AV report grants from Metabiota Inc, outside the submitted work.

269

270 **Funding**

271 This work was supported by the National Institutes of Health R01-AI139761.

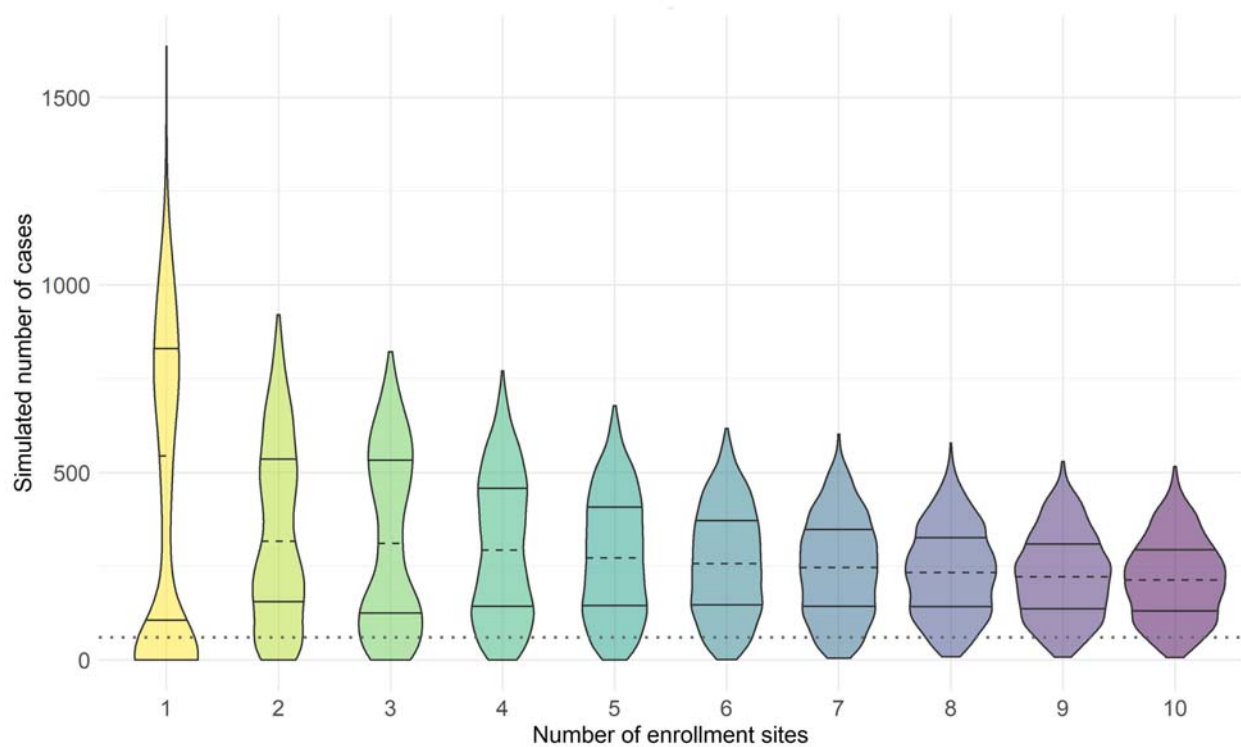
272 **References**

- 273 1. Widdowson M-A. Implementing an Ebola vaccine study—Sierra leone. *MMWR supplements*
274 2016; 65.
- 275 2. Dean NE, Gsell P-S, Brookmeyer R, et al. Design of vaccine efficacy trials during public health
276 emergencies. *Science Translational Medicine* 2019; 11: eaat0360.
- 277 3. Dean NE, Gsell P-S, Brookmeyer R, et al. Creating a framework for conducting randomized
278 clinical trials during disease outbreaks. Mass Medical Soc, 2020.
- 279 4. Dean NE, Pastore y Piontti A, Madewell ZJ, et al. Ensemble forecast modeling for the design of
280 COVID-19 vaccine efficacy trials. *Vaccine* 2020; 38: 7213-7216. DOI:
281 10.1016/j.vaccine.2020.09.031.
- 282 5. Halloran ME, Auranen K, Baird S, et al. Simulations for designing and interpreting intervention
283 trials in infectious diseases. *BMC Medicine* 2017; 15: 1-8.
- 284 6. Zhang Q, Sun K, Chinazzi M, et al. Spread of Zika virus in the Americas. *Proceedings of the*
285 *National Academy of Sciences* 2017; 114: E4334-E4343.
- 286 7. Asher J, Barker C, Chen G, et al. Preliminary results of models to predict areas in the Americas
287 with increased likelihood of Zika virus transmission in 2017. *BioRxiv* 2017: 187591.
- 288 8. Vannice KS, Cassetti MC, Eisinger RW, et al. Demonstrating vaccine effectiveness during a
289 waning epidemic: A WHO/NIH meeting report on approaches to development and licensure of
290 Zika vaccine candidates. *Vaccine* 2019; 37: 863-868.
- 291 9. Balcan D, Gonçalves B, Hu H, et al. Modeling the spatial spread of infectious diseases: The
292 GLobal Epidemic and Mobility computational model. *Journal of Computational Science* 2010; 1:
293 132-145.
- 294 10. Duffy MR, Chen TH, Hancock WT, et al. Zika virus outbreak on Yap Island, Federated States of
295 Micronesia. *The New England Journal of Medicine* 2009; 360: 2536-2543. 2009/06/12. DOI:
296 10.1056/NEJMoa0805715.

- 297 11. Bertram MY, Lauer JA, De Joncheere K, et al. Cost-effectiveness thresholds: pros and cons. Bull
298 World Health Organ 2016; 94: 925.

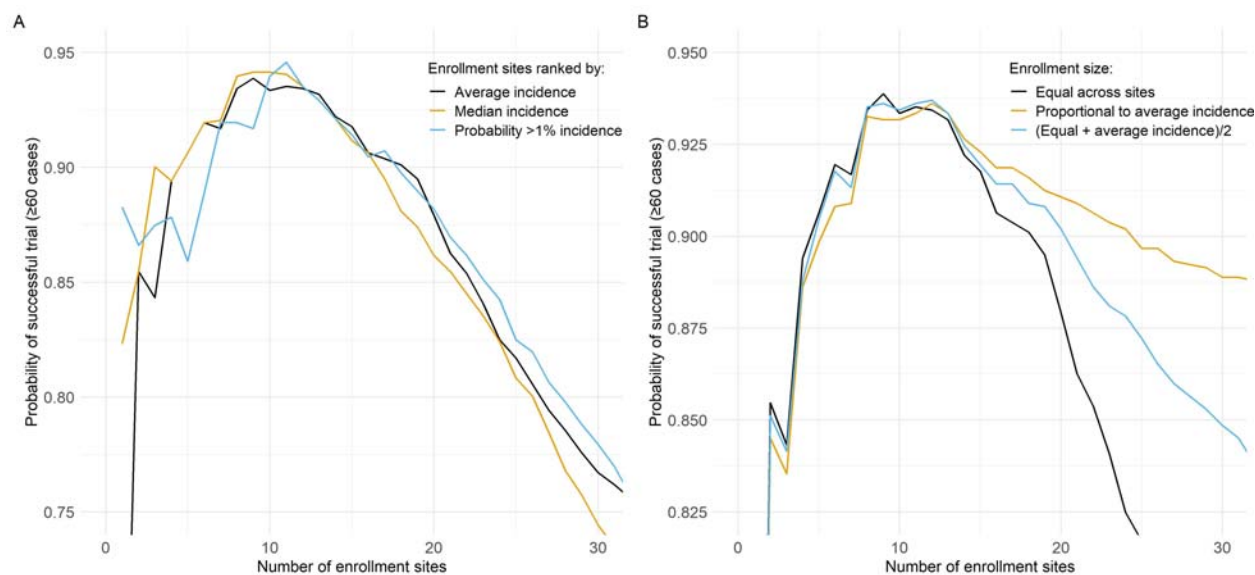
299 **Figures**

300 **Figure 1.** Violin plot of the simulated number of Zika virus cases for the top 1-10 sites with the highest
301 average site-level incidence of infection across all simulated outbreaks in one year (2017). We assume an
302 enrolled population of 15,000 across all enrollment sites with enrollment size spread evenly across all
303 sites. Median number of cases (dashed line), 25th and 75th percentiles (solid lines) are shown. The
304 threshold for a successful trial, defined as ≥ 60 cases across all sites in one year, is indicated by the dotted
305 line.



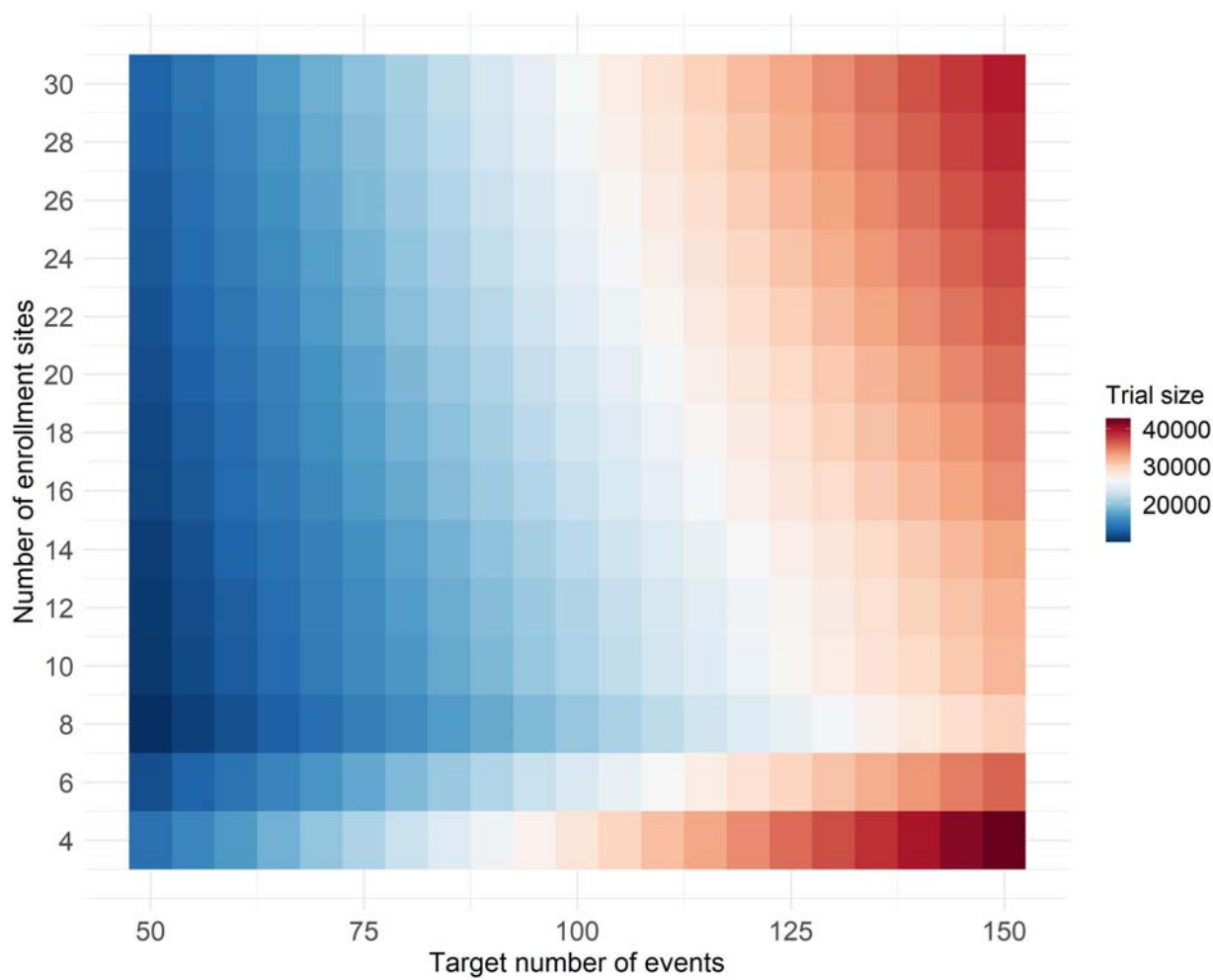
306

307 **Figure 2.** Probability of a successful trial (defined as ≥ 60 cases for an enrolled population of 15,000
308 across all enrollment sites in one year [2017]) as function of the cumulative number of enrollment sites. In
309 Panel A, enrollment size was spread evenly across all sites and sites were added sequentially based on
310 their ranking by the 1) average site-level incidence, 2) median incidence, and 3) probability of exceeding
311 1% site-level incidence of infection across all simulated outbreaks. In Panel B, sites were added
312 sequentially based on their ranking of average site-level incidence of infection across all simulated
313 outbreaks and enrollment size was 1) spread evenly across all sites, 2) proportional to the average site-
314 level incidence of each site, and 3) average of equal enrollment and proportional to mean incidence.



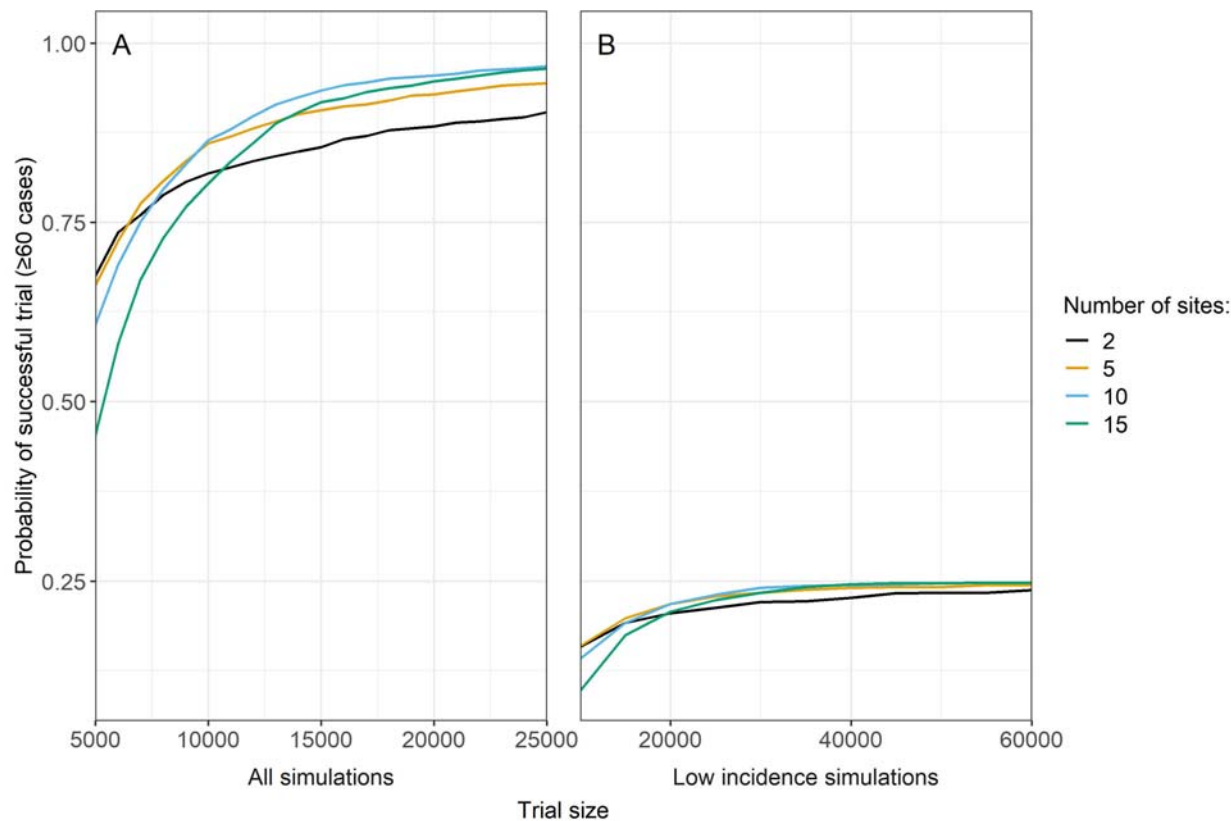
315

316 **Figure 3.** The minimum enrollment size required for an individually randomized vaccine efficacy trial to
317 achieve at least 90% probability of success for different target numbers of events and numbers of sites.
318 Sites were ranked by average site-level incidence of infection across all simulated outbreaks in one year
319 (2017) with enrollment proportional to average site-level incidence of infection.



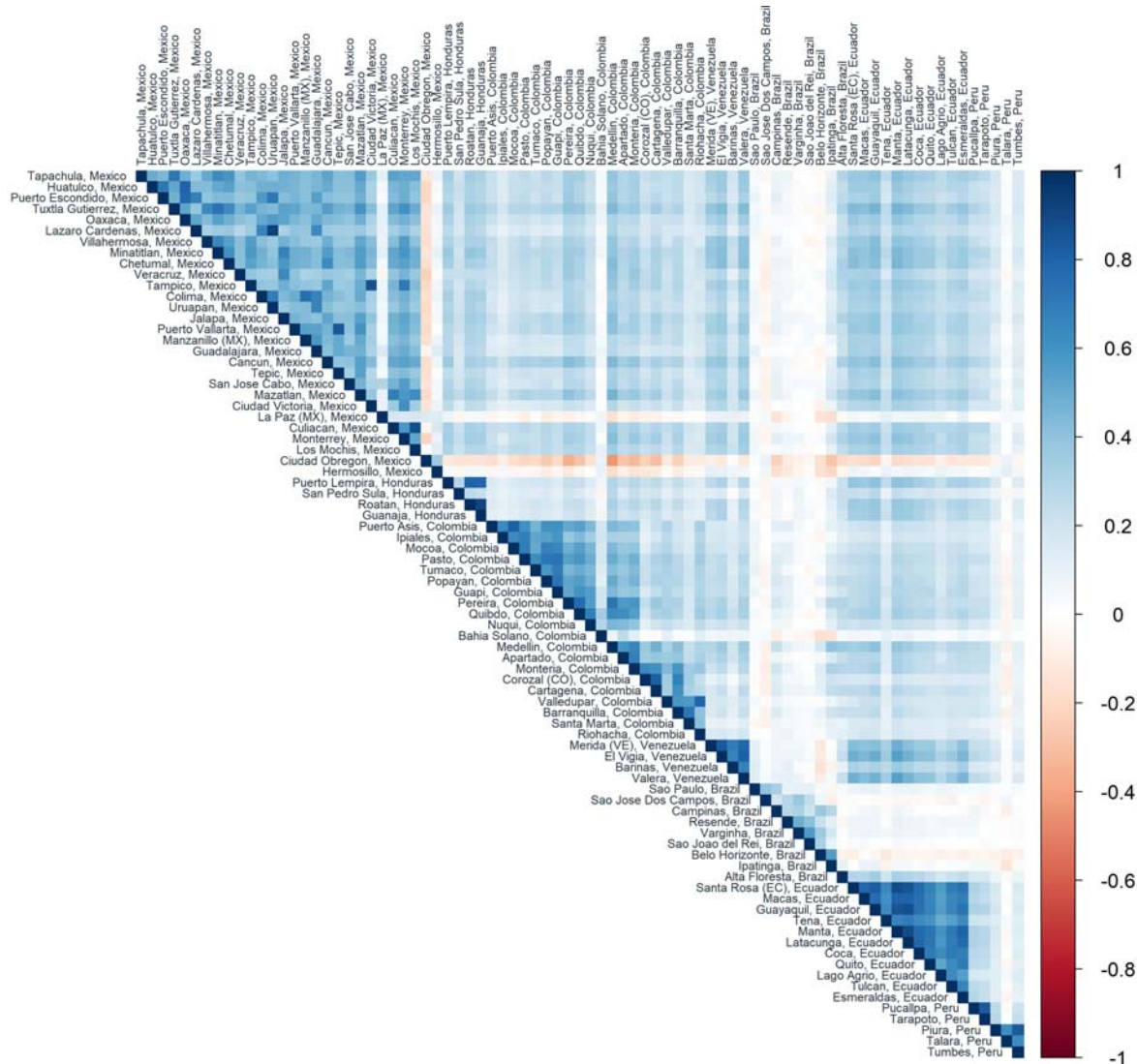
320

321 **Figure 4.** Probability of a successful trial (defined as ≥ 60 cases) as function of trial size, with enrollment
322 size spread evenly across all sites. Sites were added sequentially based on their ranking by average site-
323 level incidence of infection across all simulated outbreaks in one year (2017). Panel A includes all
324 simulations, whereas Panel B is restricted to the 25% of simulated epidemics with lowest overall
325 incidence across all sites.



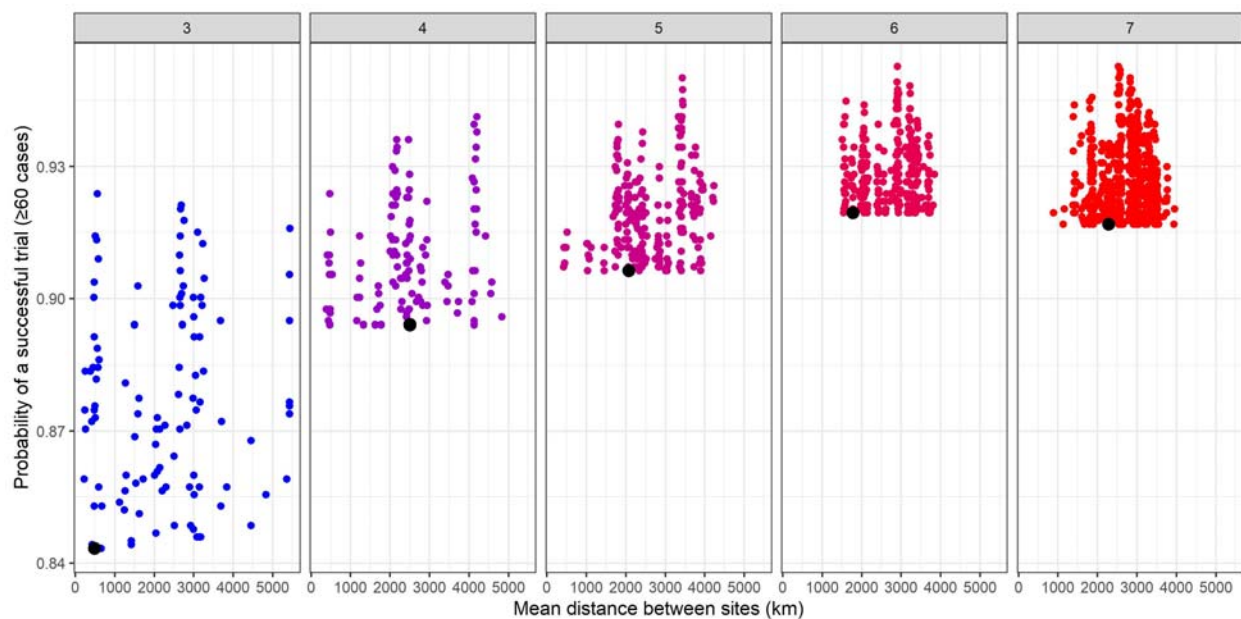
326

327 **Figure 5.** Spearman correlation of average site-level incidence of infection across all simulated outbreaks
328 in one year (2017) between sites identified by the global epidemic and mobility model (GLEAM)
329 (countries with at least four sites included). Sites are sorted by country and then by latitude.



330

331 **Figure 6.** Probability of a successful trial (defined as ≥ 60 cases for an enrolled population of 15,000
332 across all enrollment sites in one year) by the mean distance between sites (in kilometers). The panels
333 represent numbers of sites and points represent combinations of sites from the top 15 sites with the
334 highest average site-level incidence of infection across all simulated outbreaks in one year (2017) that had
335 a higher probability of success than the combination of sites with the highest projected incidence
336 (represented by the black dot). Enrollment is assumed to be spread evenly across sites. We have included
337 combinations of 3-7 sites of the top 15 sites for illustration, but this process could be extended to any
338 number of sites.



339