THE LANCET Infectious Diseases

Supplementary webappendix

This webappendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Bellan SE, Pulliam JRC, Pearson CAB, et al. Statistical power and validity of Ebola vaccine trials in Sierra Leone: a simulation study of trial design and analysis. *Lancet Infect Dis* 2015; published online April 15. http://dx.doi.org/10.1016/S1473-3099(15)70139-8.

The statistical power and validity of Ebola vaccine trials in Sierra Leone: A simulation study of trial design and analysis

Steven E. Bellan, Juliet R. C. Pulliam, Carl A. B. Pearson,
David Champredon, Spencer J. Fox, Laura Skrip, Alison P. Galvani, Manoj Gambhir, Ben A. Lopman, Travis C. Porco, Lauren Ancel Meyers, Jonathan Dushoff

Online Supplementary Appendix

All 'R' scripts used to perform the analyses in this manuscript and its supplementary material are provided in the following public GitHub repository: https://github.com/sbellan61/EbolaVaccPowerSL.git.

Validity of a Study Design and the False Positive Rate

The validity of a study hinges on the absence of random error, selection and information bias, confounding, and reverse causality as factors influencing the study findings. Here, we focus on the first of these criteria, highlighting when studies are likely to underestimate the effect of random error and, consequently, overestimate the precision of their results. Throughout the main text, we use the term 'validity' to indicate that a study exhibits a false positive rate, or alpha level, at or below the pre-specified target. We note that pre-specification of an alpha level (usually at 0.05) is fundamental to all frequentist hypothesis-testing frameworks, and also underlies the coverage of confidence intervals (commonly 95%). By specifying the false positive rate at 0.05, investigators are willing to erroneously conclude that there is an effect (when there is none) 5% of the time. Similarly, when investigators use 95% confidence intervals, they intend that these intervals will include the true value being estimated 95% of the time (if this study were to be repeated). When these pre-specifications are made, but actual false-positive rates are higher or confidence interval coverage is lower (95% confidence intervals include the true value less than 95% of the time), a study is less conservative than intended. In these cases, the calculated P value is not a true P value, and the calculated 95% confidence interval is not a true 95% confidence interval.

Furthermore, pre-specification of the false positive rate determines how conservative an analysis is intended to be. Thus, it is not acceptable to conduct a study with a pre-specified false-positive rate of 0.05 that actually winds up having a false positive rate of 0.1 (as found in this study to be the case for standard approaches to estimating the SWCT under consideration). In analyses performing multiple testing (which also suffer from alpha inflation) certain adjustments are made to account for the increased false-positive rate induced by examining multiple hypotheses, because the inflation properties are easily calculated for this case. However, in our case and more generally, inflation of the false positive rate may not be stable across all study outcomes. That is, we found that the false positive rate increased with increasing incidence within the study population and variation in underlying hazard. In other words, simple adjustments, such as halving the alpha criterion in the hope that the false positive rate remains under 0.05 across all scenarios, cannot be justified.

On the other hand, if a 10% false positive rate were considered acceptable, then that should be the specification. Furthermore, if investigators consider an SWCT with a pre-specified 10% false positive rate, then the power of this design must be compared to that of an RCT with a 10% false positive rate. Our analysis shows that traditional analyses comparing the two trial designs are unjustifiable because the SWCT is less conservative, showing an inflated false-positive rate.

We identify an analysis that maintains the level of conservativeness between designs and then compare their power.

Explanation of Inflated False Positive Rates

Parametric regression models exhibited elevated false-positive rates because they fail to take account for the complexity of the covariance (i.e. dependency) in the data. While both the Cox proportional hazard gamma frailty (CoxPH) and Poisson regression models accounted for cluster-level variation and temporal trends, they do not account for differences in trends between clusters. This causes a mis-specification of the covariance matrices fitted by these models (analogous to analyzing cluster-randomized data without accounting for the cluster covariance structure[1]) which leads to overestimates of precision. Thus, efficacy estimates that are actually different from zero just due to noise alone, are misinterpreted as being significantly different from zero, more often than expected by the target false positive rate. This causes a bias away from the null hypothesis in both directions. Therefore, the point estimate is unbiased because it is centered on zero. But the estimated P value and confidence intervals are biased in a non-conservative direction.

Furthermore, Figure S3 illustrates that false positive rates continue to increase as the proportion of district-level cases in the trial increases up to 30%. We do not consider such high values in the main text because they seem implausible in the context of the Sierra Leone trail. However, we note that such large numbers of cases in a trial may not have been implausible at the height of this epidemic. An SWCT design at that time could have consequently exhibited false-positive rates as high as 0.15, if analyzed with standard statistical approaches and clusters exhibited greatly different trends in incidence simulated here. This is a broadly-applicable finding, with implications well beyond the Ebola vaccine trials considered here.

The bootstrapping confidence intervals exhibit inflated false-positive rates for a different reason. Bootstrapping methods are asymptotically unbiased, meaning that they work well for large sample sizes. We find that, for the sample sizes achievable in these trials, cluster-level bootstrapping is likely to also lead to elevated false-positive rates, through overestimates of precision, but primarily for small sample sizes (Figure S7).

Sensitivity Analysis to SWCT Inclusion of Person-Time

We considered three different criteria for the inclusion of person-time from an SWCT design (Figure S2). The criteria highlighted in the main text exclude person-time during the protective delay (after vaccination and before protection) but include all other person-time. The second approach additionally excluded person-time at the beginning and end of the trial, when only unvaccinated or protected person-time are observed, respectively. This approach is motivated by the assertion that person-time should not be analyzed when it cannot be compared between treatment arms in the same secular time period. However, we do not take this approach in the main text because the SWCT design can also be considered intended to make before-and-after comparisons, and therefore person-time before the trial begins can give information on the expected trends within all clusters. Nonetheless, we caution inclusion of this person-time without carefully considering the trade-offs between bias and increased power. Finally, we also consider an approach that includes all person-time in the trial starting from when the first cluster becomes vaccinated. This provides modest increases in power, but also assumes that the onset of protection is known and identical across individuals, which is unlikely to hold in most realistic contexts.

Importantly, our conclusions hold, regardless of the inclusion criteria used (Figures S3-S4). The second approach reduces power substantially, however. For that reason, the inflation of the false positive rate is not apparent when hazard is low (i.e. district-level proportion of cases in the trial). Still, when hazard is increased beyond the values considered in the main text, analyses relying on these criteria also exhibit inflated false-positive rates (Figure S3).

Forecasts

We acquired district-level EVD incidence data from the Humanitarian Data Exchange (HDX 2015), which aggregates situation reports created by the Sierra Leone Ministry of Health and Sanitation. Using these time series, we fit exponential decay models to each district, using maximum likelihood estimation assuming a negative binomial observation model for cases. We fit both the exponential decay rate and the negative binomial overdispersion parameter for each district, and then took the average of the overdispersion parameters across districts to use in forecasting (negative binomial size parameter average across districts was 1.2). Thus, each district-level fit corresponds to its own fitted exponential decay rate and a country-level overdispersion parameter that governs the amount of clustering in case data. This procedure created projections that appear (qualitatively) in-line with incidence trends to date (Figure 1, S11).

Comparison of Trial Designs

Even absent the spatiotemporal variation simulated in this study, an RCT exhibits greater power than an SWCT for several reasons. First, individual-level randomization is more powerful than cluster-level randomization because randomization at the cluster level leaves similarities between individuals and groups, decreasing the effective sample size.[1] Second, designs with balanced person-time in both study arms optimize power.[2] An RCT exhibits an even balance throughout the trial. In contrast, at the beginning of an SWCT the ratio is skewed towards more unprotected person-time, and this gradually reverses over the course of the trial. Third, each trial design forces constraints on which person-time can be included in the trial. For instance, the SWCT includes person-time before vaccination, while the RCT does not; in contrast to the previous two effects, this increases the relative power of the SWCT.

In our analysis we focus on trial designs that were considered for Sierra Leone. The SWCT was suggested because of its potential ethical advantages over individual randomization. The SWCT is a form of cluster-randomized controlled trial (CRCT). CRCT designs are often used to detect community-level effects. For instance, CRCT designs are commonly used to evaluate vaccine effectiveness, inclusive of the indirect protective benefits of vaccination (i.e. herd immunity). However, we reiterate that in this case the SWCT was chosen, not to detect the indirect benefits of vaccination, but because of its expected ethical advantages. Standard CRCT designs (e.g. parallel arm design in which not all clusters receive the intervention and ordering of intervention rollout is not random) for vaccine efficacy evaluation were never considered, to our knowledge, in this setting or anywhere else in the West African epidemic because they do not have the ethical advantages anticipated for an SWCT. Furthermore, we argue that the indirect benefits of vaccination and health care worker populations are likely to be negligible because health care workers do not contribute greatly to transmission. For these reason, we did not consider other CRCT designs in this manuscript.

Finally, we did not examine the ring vaccination trial (RVT) both because it has not been considered for Sierra Leone and because it would require a far more detailed transmission model than employed here. We do, however, note that the RVT may be well powered relative to

other designs because it targets the highest risk individuals — those who are contacts of EVD cases. However, the RVT may suffer from other pitfalls. For instance, withholding vaccines from individuals at greatest risk may lead to questionable equipoise for this trial design. Traditional statistical analyses of the RVT likely also suffer from inflated false positive rates because, like the SWCT, the RVT is a form of cluster-randomized cross-over design.

Trial Simulation

Choice of cluster and trial size was dictated by the CDC's initial SWCT plans and new phased-rollout RCT design.[3,4] Rather than vary trial population size, we considered scenarios that varied the expected proportions of district-level cases occurring in the trial population. We also only considered on cluster size. Trial populations composed of greater numbers of smaller clusters (while holding the total trial population size constant) would exhibit properties approaching those of individually randomized designs (RCTs) both in terms of power and validity as cluster size approaches 1.

Each trial cluster was considered to exist within a specific district in Sierra Leone. Because there are only 14 districts in Sierra Leone, we assumed that each cluster corresponded to a district, where districts were sampled with replacement. For each cluster, we simulated a stochastic district-level forecast for the next six months as described in the main text and above. Clusters corresponding to the same district still exhibited different forecasts due to random variation in the negative binomial projection, which could for example correspond to within-district regional variation in incidence. The mean hazard experience by a cluster in a given week was then defined as:

$$\frac{(\#\text{ of district-level cases projected in a week}) \times (\text{proportion of cases expected to occur in cluster})}{300\ \text{individuals} \times 1\ \text{week}},$$

where the proportion of district-level cases that were expected to occur in the cluster, in the absence of vaccination, was varied between 2.5%, 5%, 7.5%, and 10%. For brevity, elsewhere in the text we state that these percentages correspond to the assumed district-level proportion of cases that occurred in the trial. However, more accurately, these percentages correspond to the percentage of district level cases that would occur in a given cluster. For example, if two clusters occur in a district (due to the resampling with replacement procedure above) at the 5% assumption, this corresponds to 10% of district-level cases occurring within those two clusters combined.

Individuals became infected as dictated by exponentially distributed waiting times. Each week, each individual's hazard was calculated as the product of the mean cluster hazard, their individual risk deviate (Figure 3), and, for individuals who were vaccinated and past their seroconversion delay, (1 - vaccine efficacy). Starting from the trial start date, we drew an exponentially distributed time until infection for every individual for each week using these individual-week-specific hazards, and censoring infection times greater than a week (which were interpreted to mean that the individual did not get infected that week and was therefore at risk the following week). This process was repeated each week up until an individual became infected, or until the end of the trial (24 weeks), thus relying on the memoryless nature of the exponential waiting time distribution.

Trial Analysis

We performed 2000 simulations for each parameter set. The number of simulations run was chosen to provide precise estimates of the false positive rate. Nominally set false positive rates of 5% would yield 100 simulations giving false positives out of 2000. Based on the binomial proportion variance, p(1-p)/N, this would yield false positive rate estimates of 5% +/- 1%, and far more precise estimates of power.

We assumed a 21 day delay between vaccination and development of protective immunity[4–6] (hereafter, denoted 'protective delay') corresponding to the two potential candidate vaccines, rVSV and ChAd3, but conducted a sensitivity analysis considering a 5 day delay to evaluate the scenario in which a vaccine provides post exposure prophylaxis, such as might be the case for rVSV.[5]

We assume that the protective delay is unknown and, consequently, exclude person-time during that delay from all analyses (except in the sensitivity analysis described above). Including this person-time in an SWCT as unprotected person-time increases power (Figure S4). While information about the distribution of this delay may be provided by immunogenicity studies (e.g., during Phase I/II trials), the relationship between immunological assay reactivity and protective immunity is poorly understood. Furthermore, variation among individuals in the duration of the protective delay complicates characterization of this person-time in real-world scenarios. Thus, we chose to exclude this person-time, as has been done in other trial protocols.[4]

We did not consider adaptive designs, in which the data are analyzed at interim checkpoints and the trial is stopped once efficacy is definitively determined. Ending a trial early requires interim evaluations of vaccine efficacy (i.e. sequential designs). Interim evaluations require downward adjustment of the alpha level at each interim check to maintain the overall target rate of 0.05.[7] Because interim analyses come at the cost of statistical power[7] and sufficient power is a major concern for the proposed study,[8] we assumed that the data are analyzed only once, at six months. However, future work is warranted to evaluate the tradeoffs between trial power and speed during a declining epidemic.

We considered many parametric regression models for analysis of simulated trial data including Cox proportional hazards gamma frailty models (CoxPH), generalized linear models with cluster-level fixed effects (GLMF), generalized linear mixed models (GLMM), and generalized estimating equations (GEE) with autoregressive (AR1) covariance-variance matrices. For the latter three models, we aggregated the number of infections amongst trial participants by cluster and by week, and analyzed the resulting data with Poisson regression that included the log observed person-time as an offset term and time as a log-linear predictor variable,

$$Y_{i,v,t} \sim \text{Poisson}(\lambda_{i,t})$$
$$log(\lambda_{i,v,t}) = \beta_{vacc} X_{i,v,t} + log(U_{i,v,t}) + \beta_{time} t + Z_i$$

where $Y_{i,v,t}$ corresponds to the number of cases observed in cluster i with vaccine treatment assignment v at time t, $\lambda_{i,v,t}$ corresponds to the hazard of those individuals, β_{vacc}corresponds to the log(relative hazard) of infection amongst vaccinated and protected versus unprotected individuals, $X_{i,v,t}$ is an indicator variable for vaccine status, $U_{i,v,t}$is the amount of person-time observed during time interval [t, t+1), β_{time} is the log(relative hazard) for each one unit increase in time, and Z_i is a cluster-level, normally distributed, random effect. GEE models included autoregressive covariance structure in Z_i. For GLMF models, Z_i were considered fixed effects.

In preliminary analyses, we found that the GLMM and GEE models were very unstable for small sample sizes and often could not be fitted to all trial simulations (i.e. because they were divergent). However, even in these preliminary simulations, we found that the GEE performed very poorly (i.e. exhibited a very high false positive rate) and the GLMM performed comparably to the GLMF, despite the increased degrees of freedom spent by the latter. Because (1) bootstrap- and permutation test-based inference only rely on point estimates of vaccine efficacy, not precision estimates (i.e. P values or confidence intervals), (2) GLMM and GLMF exhibit approximately the same point estimates, and (3) GLMM and GEE models were divergent for many simulations, we only simulated the more stable GLMF estimator for our power analyses.

Even after excluding GLMM and GEE models, we found CoxPH and GLMF models were still ill-adapted for rare events. These models provide infinite confidence intervals when zero cases occur in one of the treatment groups, as would occur for instance with very low infection risk (e.g. the scenario with 2.5% proportion of district-level cases occurring in the trial) and a relatively effective vaccine. Certain shrinkage estimators exist to solve these problems, but are model-specific.[9] To facilitate variance and confidence interval calculation for all models from a common correction, we took an approach similar to that commonly used by the Mantel-Hanzel odds ratio estimator, which adds 0.5 to all cells in a contingency table to avoid problems associated with zero cells. More precisely, if zero cases occurred in one of the treatment arms of a simulated trial (e.g. 18 control cases, 0 vaccinated), we selected the uninfected individual at greatest hazard of infection in each of the two arms and assumed they were infected during the week when their hazard was highest (e.g. yielding 19 control cases, 1 vaccinated), to provide conservative lower confidence interval bounds on vaccine efficacy.

Each regression framework, when applied to a simulated data set, yielded an estimated vaccine efficacy, $V = 1 - exp(\beta_{vacc})$ with associated parametric confidence intervals based on the model specification.

We then used these same regression estimates to perform two nonparametric analyses, cluster-level bootstrapping and a permutation test. Bootstrapping methods estimate the variance of an (efficacy) estimate by resampling the observed data with replacement at the individual or cluster level to estimate the variance or confidence intervals of the estimate. Permutation tests construct a null hypothesis from the observed data (in this case that the order of vaccination was not associated with infection) and then scramble (i.e. permute) the vaccination assignment of individuals (RCT) or clusters (SWCT) to generate permuted estimates. These permuted estimates can then be used to quantify the probability of observing an estimate as or more extreme than the observed estimate if the vaccine indeed had no effect (i.e. a P value).

Cluster-bootstrap

Within the bootstrap framework, vaccine efficacy is established when the bootstrap confidence intervals around efficacy estimates exclude the null value (zero efficacy). To generate bootstrap confidence intervals, we resampled the 20 simulated clusters with replacement,[10] and then calculated an estimated vaccine efficacy from the bootstrap sample of clusters, V_b. This procedure was repeated 200 times and the 2.5% and 97.5% quantiles of the empirical distribution of bootstrap efficacy estimates $\{V_{b,1}, \ldots, V_{b,200}\}$ were used as the bootstrap confidence intervals. When analyzing real empirical data, more bootstrap samples are advised. However, we chose fewer resamples for each simulation due to the factorial computational burden associated with using resampling methods over so many simulations and because we average results over 2000 simulations for each of 300 parameter sets (including all sensitivity

analyses). Across all 600,000 simulations, doing 200 bootstrap samples in 200 permutation samples, we fit each regression model 600,000*400 = 240 million times. Because cluster-level bootstrapping is only relevant for cluster-randomized designs, we did not generate bootstrap estimates from RCT simulations (and did not consider individual-level bootstrapping because parametric models for the RCT proved valid). Contiguous block bootstrap methods (i.e. sampling with replacement contiguous portions of cluster time series) are recommended for spatiotemporally heterogeneous data. However, these methods require choosing contiguous block size (i.e. length of contiguous time periods resampled from the data) based on information in the autocorrelation in the observed data, which is generally problematic[11] and, in particular, likely to be difficult with such sparse data as would be acquired by the proposed trials.

Permutation test

Permutation tests construct empirical distributions under the null hypothesis relying only on the available data and, because they make no parametric assumptions and do not rely on asymptotic theory (unlike the bootstrap), are always valid, albeit less powerful.[12,13] For the SWCT design, the null hypothesis that the vaccine has no efficacy equates to the hypothesis that vaccination timing has no effect on infection risk within a cluster. Thus, for SWCT analyses, we performed permutation-based inference by randomly permuting vaccination times of each cluster while holding the infection outcomes constant, and calculating a parametric regression-based vaccine efficacy estimate, V_p. This procedure was performed 199 times (because technically the un-permuted sample, with its estimate denoted V, is one of the permuted samples because under the null hypothesis it should be no different), and a P value was calculated by determining the proportion of these estimates $\{V, V_{p,1}, \ldots, V_{p,199}\}$ that were more extreme (in either tail) than the estimate from the non-permuted data. For the RCT, we performed permutation-based inference by permuting the treatment (i.e. vaccination vs. control) assignment of individuals within each cluster.

Generating confidence intervals with permutation tests is challenging.[14] Parametric regression provides reliable confidence intervals for efficacious vaccines for RCT but not SWCT designs, with 95% confidence interval coverage for the latter as low as 90% for a vaccine efficacy of 0.9 (Table S1). Thus, future work is needed to develop appropriate confidence intervals for the SWCT in spatiotemporally variable settings. For now, we recommend that SWCT data be analyzed by first applying permutation tests to assess the null hypothesis that the vaccine is ineffective and then, if rejected, applying a parametric CoxPH model to calculate confidence intervals for vaccine efficacy, while acknowledging that their coverage may be low.

Performing all parametric and nonparametric analyses on all 600,000 simulations took a total of 6,000 node-hours of computing time on the 12-core nodes of the Lonestar Linux Cluster of the Texas Advanced Computing Center.

Cited References

1 Cornfield J. Randomization by group: a formal analysis. *Am J Epidemiol* 1978; **108**: 100–2.

2 Kalish L a, Harrington DP. Efficiency of balanced treatment allocation for survival analysis. *Biometrics* 1988; **44**: 815–21.

3 Lopman B. Can effectivness be accurately estimated using a phased introduction of an Ebola vaccine? In: Modeling the Spread and Control of Ebola in West Africa. Atlanta, 2015. Accessed February 12, 2015 from http://bioinformatics.gatech.edu/ebola-modeling-workshop-program.

4 Centers for Disease Control and Prevention. STRIVE (Sierra Leone Trial to Introduce a Vaccine Against Ebola). In: ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US) 2015. Available from https://clinicaltrials.gov/ct2/show/NCT02378753 NLM Identifier: NCT02378753.

5 Geisbert TW, Feldmann H. Recombinant vesicular stomatitis virus-based vaccines against Ebola and marburg virus infections. *J Infect Dis* 2011; **204**. doi:10.1093/infdis/jir349.

6 Rampling T, Ewer K, Bowyer G, *et al.* A Monovalent Chimpanzee Adenovirus Ebola Vaccine — Preliminary Report. *N Engl J Med* 2015; 150202093719007.

7 Todd S. A 25-year review of sequential methodology in clinical studies. Stat. Med. 2007; **26**: 237–52.

8 Mohammadi D. Ebola vaccine trials back on track. *Lancet* 2015; **385**: 214–5.

9 Ambler G, Seaman S, Omar RZ. An evaluation of penalised survival methods for developing prognostic models with rare events. *Stat Med* 2012; **31**: 1150–61.

10 Ren S, Lai H, Tong W, Aminzadeh M, Hou X, Lai S. Nonparametric bootstrapping for hierarchical data. *J Appl Stat* 2010; **37**: 1487–98.

11 Davies MM, Laan MJ Van Der. A New Approach to Variance Estimation for Time-ordered Dependent Data. *UC Berkeley Div Biostat Work Pap Ser* 2014. Accessed on February 15 from http://biostats.bepress.com/ucbbiostat/paper322.

12 Robins JM. Confidence intervals for causal parameters. *Stat Med* 1988; **7**: 773–85.

13 LaVange LM, Durham TA, Koch GG. Randomization-based nonparametric methods for the analysis of multicentre trials. *Stat Methods Med Res* 2005; **14**: 281–301.

14 Rigdon J, Hudgens MG. Randomization inference for treatment effects on a binary outcome. *Stat Med* 2015; **34**: 924–35.

Table S1. Bias and coverage of vaccine efficacy estimators. Values shown are for a true vaccine efficacy of 0.9 and a protective delay of 21 days, and RCT results presented are for a risk-prioritized RCT. For both trial designs the bias and confidence interval coverage shown correspond to vaccine efficacy estimates and 95% confidence intervals from a Cox proportional hazards mixed effects model (CoxPH). The CoxPH model exhibits inflated false positive rates (i.e., performs poorly when a vaccine has no effect), and confidence interval coverage is also slightly low even when vaccine efficacy is high. We recommend using the permutation test to determine whether to reject and a null hypothesis of an ineffective vaccine and, if rejecting the null hypothesis, to use parametric confidence intervals from the Cox proportional hazards model (since permutation tests do not provide confidence intervals), but acknowledging that these confidence intervals are still too narrow.

	RCT		SWT	
% of district-level cases in trial	**95% CI coverage**	**bias**	**95% CI coverage**	**bias**
2.50%	0.94	0.0062	0.90	-0.03
5%	0.95	-0.0047	0.93	-0.033
7.50%	0.96	-0.0059	0.92	-0.049
10%	0.96	-0.0035	0.92	-0.039

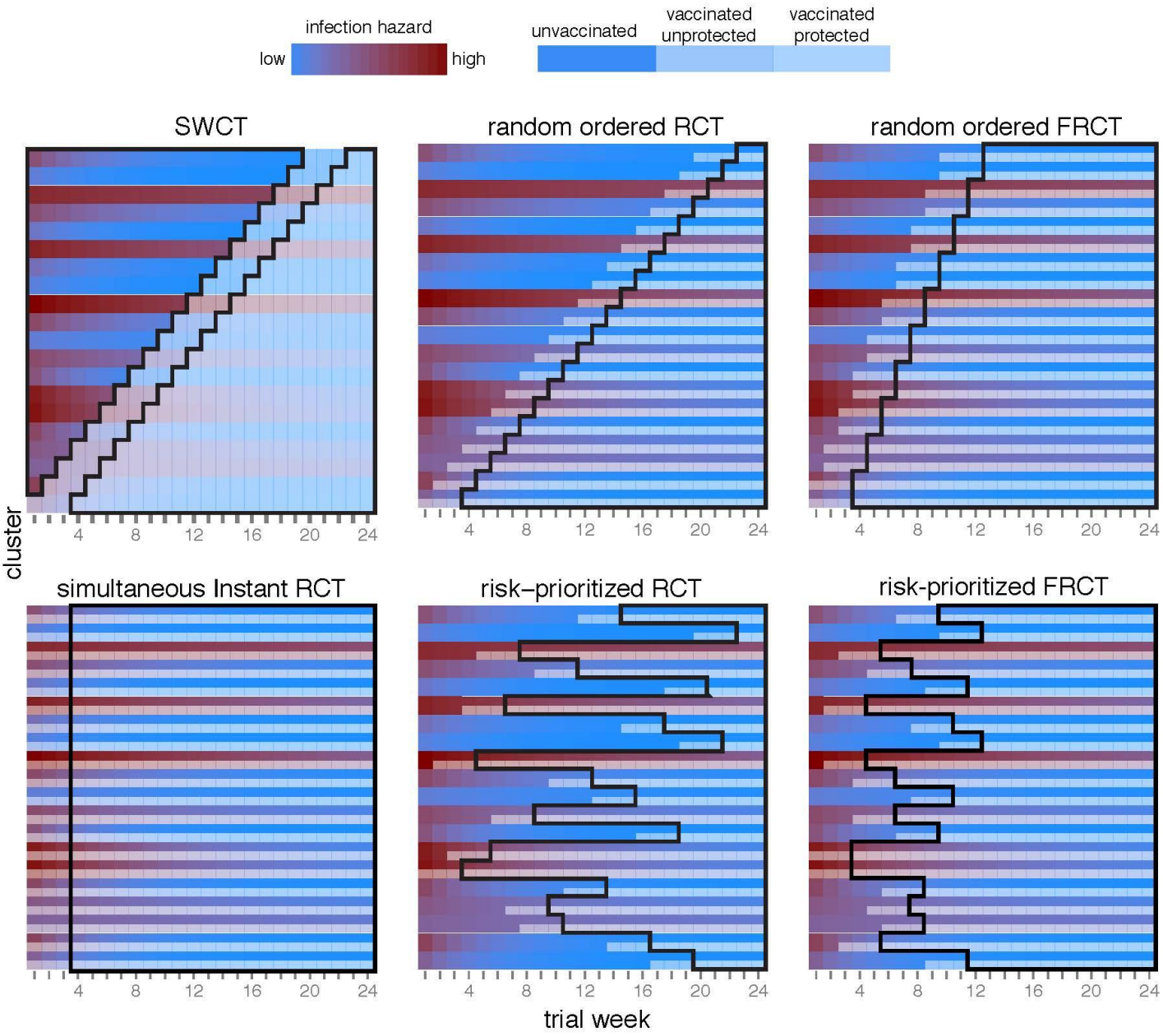

Figure S1. Schematic diagram of trial designs. Same as Figure 3 but also showing random-ordered RCT and FRCT, as well as displaying the person-time analyzed in each design (outlined in black). In an FRCT, individuals are vaccinated at the same rate as in the SWCT by vaccinating half of each of two clusters per week. However, in contrast to the SWCT, only half the individuals in the trial become vaccinated. The risk-prioritized RCT best captures high-hazard person-time while also randomizing within high-risk clusters. The risk-prioritized FRCT and simultaneous instant RCT only capture a small amount more person time than the risk-prioritized RCT. The simultaneous instant RCT is only shown as an ideal comparator, but does not satisfy the logistical constraints adhered to by other designs.

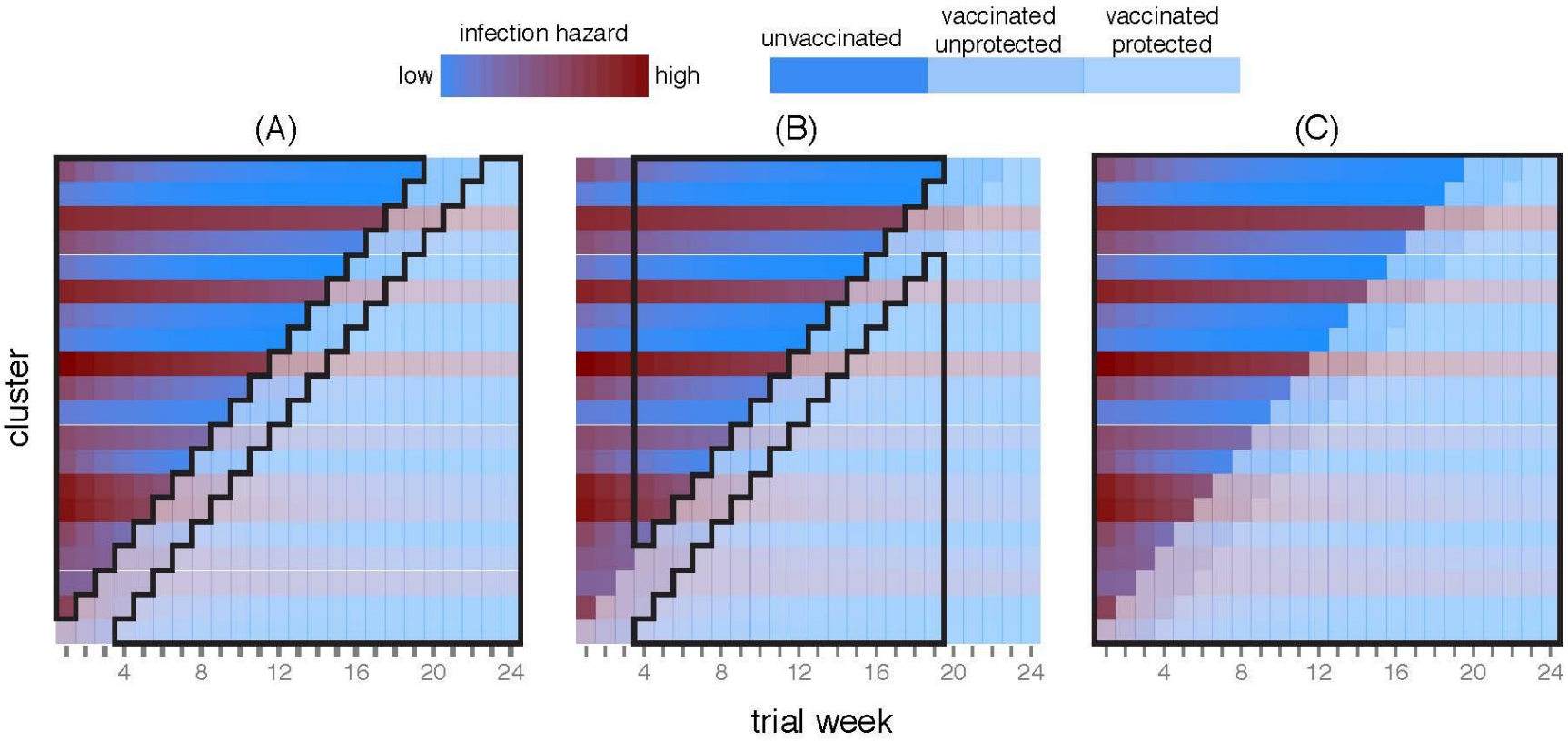

Figure S2. Schematic diagram of person-time options for stepped wedge cluster trial (SWCT) designs. In our main analyses we included all person-time in SWCT analysis except for the 21-day protective delay (A). However, we also conducted a sensitivity analysis either additionally excluding the beginning and end of the trial when only unvaccinated individuals and protected individuals are observed, respectively (B) and an analysis that includes all person-time, correctly including the protective delay as unprotected person-time (C).

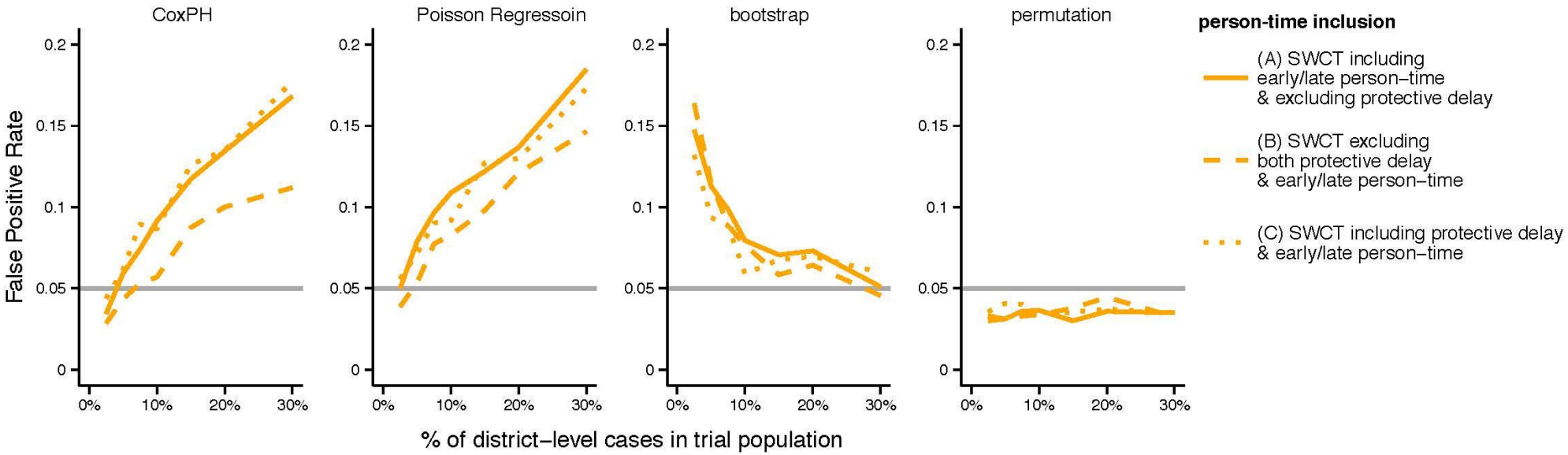

Figure S3. False positive rate for SWCT by person-time inclusion criteria. The proportion of simulations erroneously rejecting the null hypothesis that the vaccine has no effect on risk is shown by statistical method (panels), person-time inclusion criteria for the SWCT (line types, corresponding to Figure S2), and percentage of district-level cases in the trial population (X axis). Here, we show scenarios with up to 30% of district-level cases in the trial population, while only showing up to 10% in the main text because values greater than 10% may be implausible for a trial population of 6,000 in Sierra Leone at this time. We show greater values for two reasons. First, larger values correspond to greater incidence in the trial population. Incidence levels this high could have been plausible at the height of the epidemic, particularly in a much larger trial population. Second, we show greater values here to illustrate that even option (B) for person-time inclusion is invalidated by inflated false-positive rates when trial incidence is great enough. The bootstrap and permutation analyses shown here were performed over the CoxPH efficacy estimator.

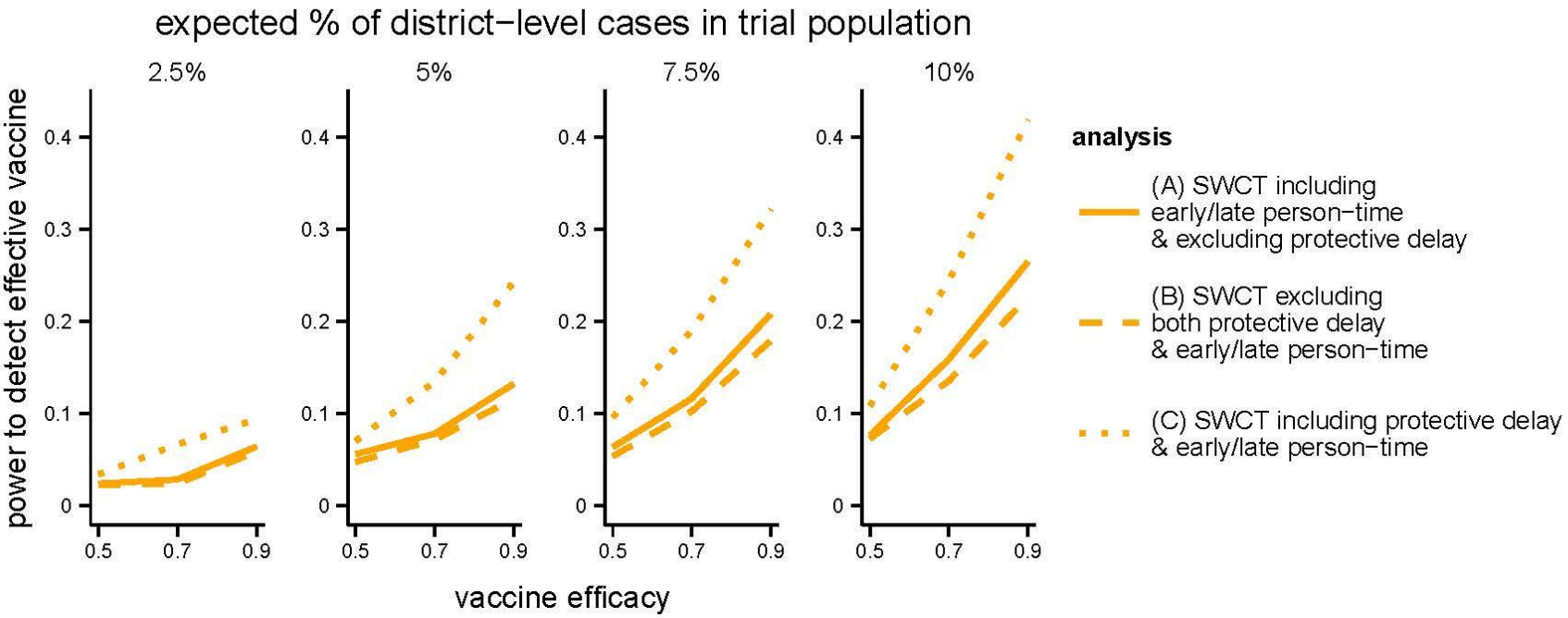

Figure S4. Power for SWCT by person-time inclusion criteria. Power to detect an effective vaccine is shown by percentage of district-level cases in the trial population (panels), person-time inclusion criteria for the SWCT (line types, corresponding to Figure S2), and vaccine efficacy (X axis). Analyses shown here were performed with the permutation test approach (over the CoxPH estimator) because, unlike other analyses, it does not exhibit inflated false positive rates. Including the period between vaccination and protective immunity in the analysis (option (C)) increases power, but would require this period to be well-characterized and with little individual-level variation.

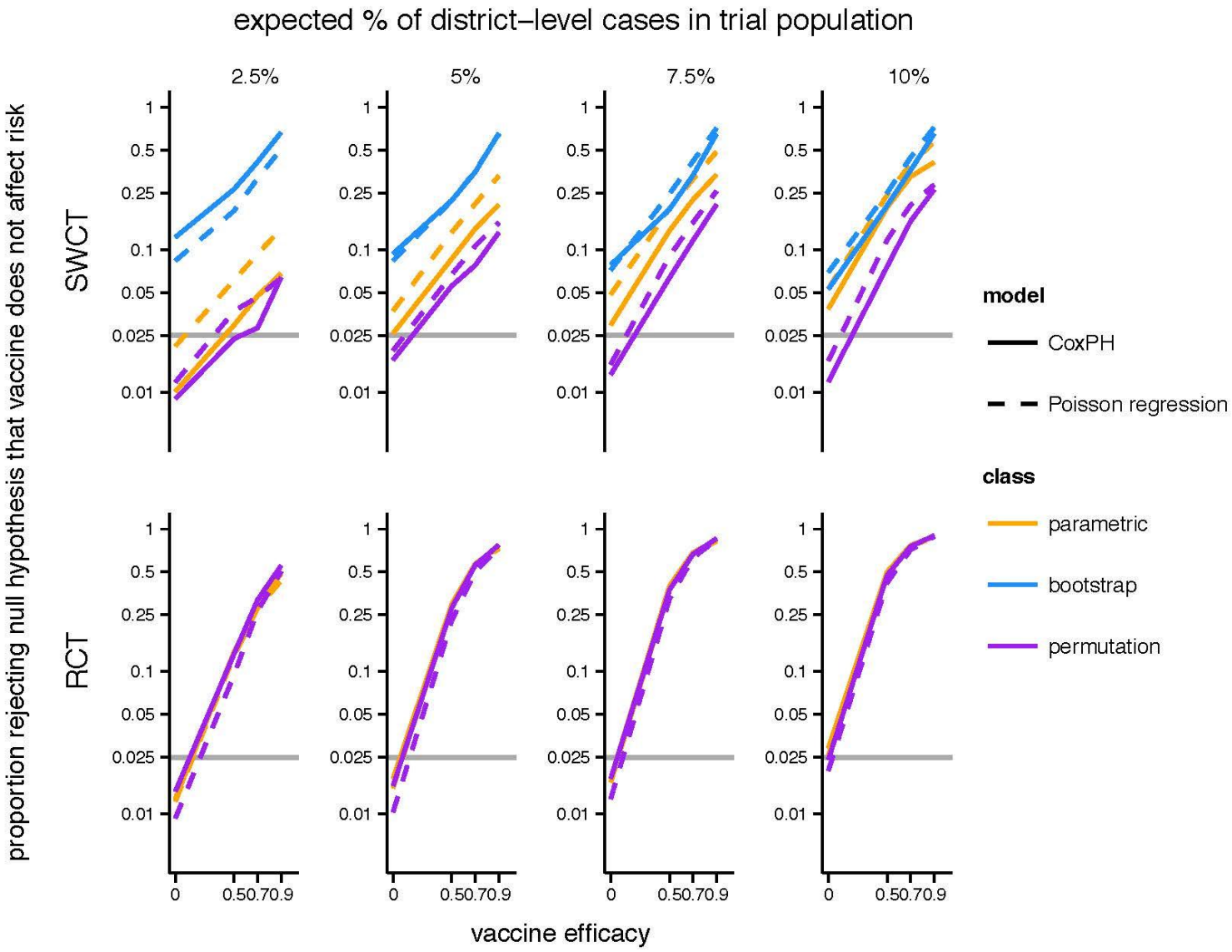

Figure S5. Power by model class. Shown for Cox proportional hazards frailty model (CoxPH) and Poisson regression model with cluster-level fixed effects where inference is performed either using variance estimates from the Cox model itself (parametric), from a cluster-level bootstrap of the Cox estimator (not performed for the RCT given individual randomization), or from a permutation test of the Cox estimator. Higher power of the bootstrap and parametric approaches comes at the cost of elevated false positive rates, shown by their Y-intercept occurring above the 0.025 one-tailed α criterion. Despite the modest increase in power from the Poisson regression over the CoxPH, we focus on the latter in the main text because it is more robust to changes in epidemic trends (the former assumes a monotonic trend in hazard over the course of the trial). While permutation tests often have lower power than parametric methods, the difference in power between these methods in assessing the SWCT (Figure S4) is artificially magnified by the difference in false positive rates (i.e. the red lines are only greater than the blue lines at positive vaccine efficacies because their false positives (y-intercepts) are shifted up). Thus, the power of SWCT and RCT designs is most appropriately assessed by comparison of permutation tests for the SWCT with any valid study design for the RCT.

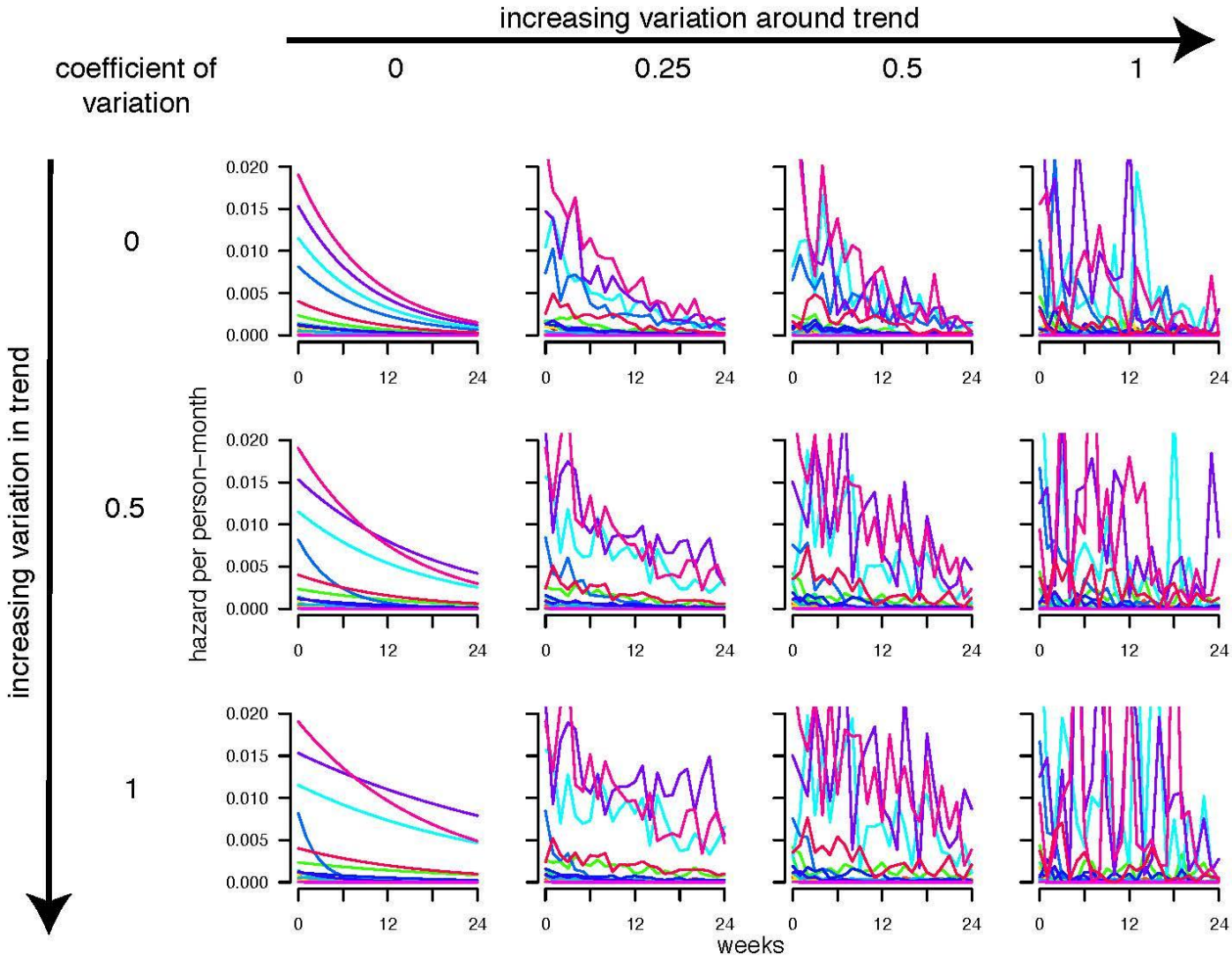

Figure S6. Simulated hazard trajectories for varying degrees of inter-cluster heterogeneity in trends. These phenomenologically simulated hazard trends were used to produce Figure S7 and explore the sensitivity of different trial designs and analytical approaches to these multiple aspects of heterogeneity. For the main text results, we simulated cluster-specific hazards from Sierra Leone district-level incidence models (Figures 1-2). Example simulated hazards for 20 clusters (each line representing a cluster) for different combinations of variation in cluster-specific hazard decay rates and degrees of temporal fluctuation. Each cluster was assigned a hazard at week 0 from a gamma distribution with mean of 0.002 and coefficient of variation 1. Each cluster's proportional weekly decay rate was drawn from a logistic-normal distribution (to require that all clusters decline with decay rate < 1) with mean 0.9 and considering coefficients of variation of 0, 0.5, and 1. Thus, the mean hazard in cluster *c* in week *t* is given by $\lambda_{c,t} = \lambda_{c,0} \times {r_c}^t$ where $\lambda_{c,0}$is the hazard in that cluster at the start of the trial and r_c is the logistic-normally distributed decay rate. Finally, the temporal fluctuations or noisiness of each cluster's hazard trajectories was determined by sampling a gamma distributed variable around the smooth exponential decay trend with mean equal to the smooth trend and considering coefficients of variation of 0, 0.25, 0.5, and 1.

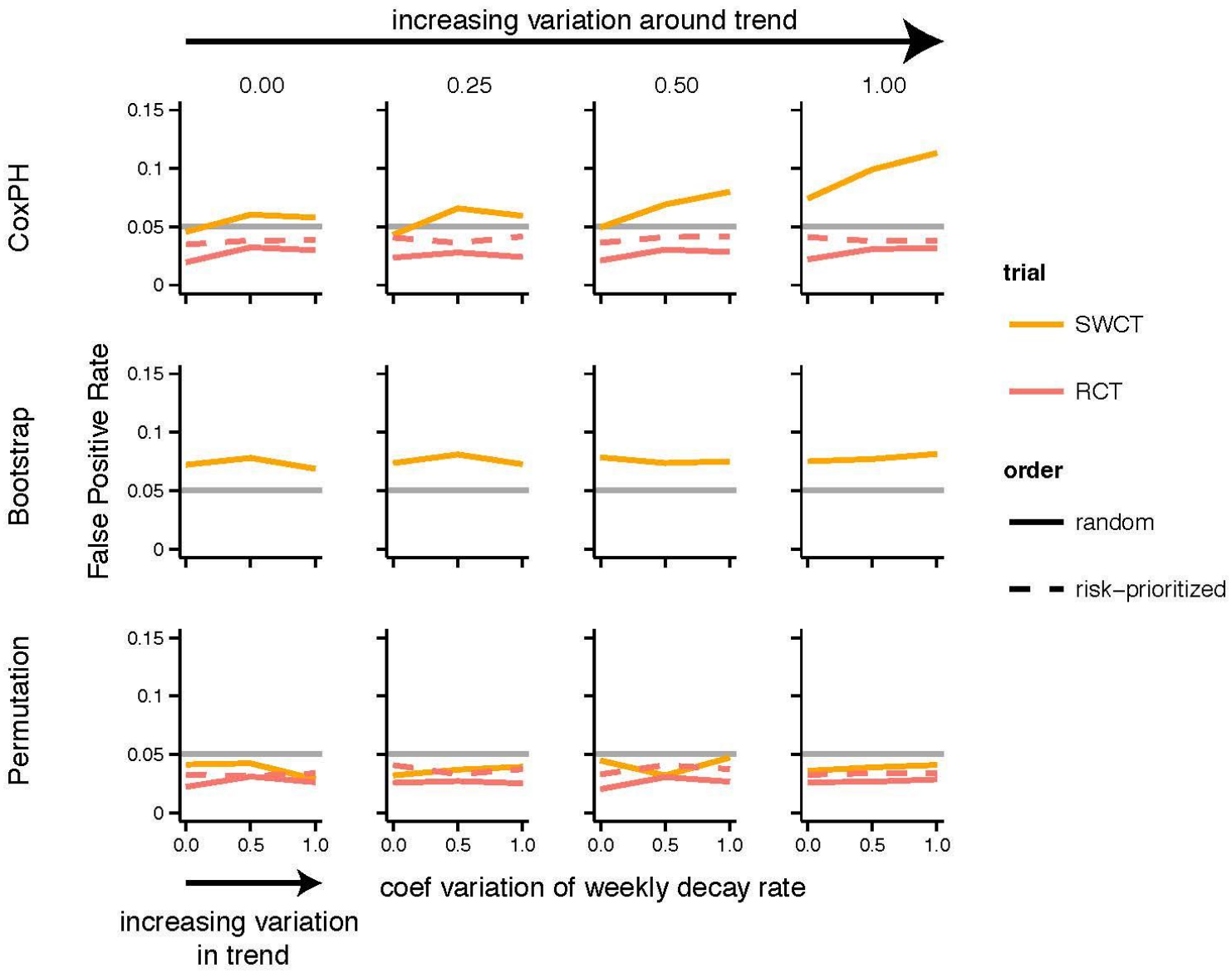

Figure S7. False positive rates for varying degrees of inter- and intra-cluster heterogeneity in trends. False positive rates for Cox proportional hazards gamma frailty models (CoxPH; first row), cluster-bootstrapping on CoxPH estimates (second row), and permutation tests on CoxPH estimates (third row), for varying degrees of variation in the weekly decay rate (X axis in each panel, corresponding to each row in Figure S6) and varying degrees of fluctuation around a smooth trend (columns, corresponding to columns in Figure S6). CoxPH models of SWCT data perform well with decreasing hazards and cluster-level variation in baseline hazard, but exhibit elevated false positive rates when the rates of decrease are different between clusters or for noisier temporal trends, with a strong interaction between these two effects. Cluster-bootstrapping exhibits inflated false positive rates regardless of heterogeneity. All analyses of RCTs and all permutation test approaches exhibit valid false positive rates. Bootstrap and permutation approaches were performed over the CoxPH estimator.

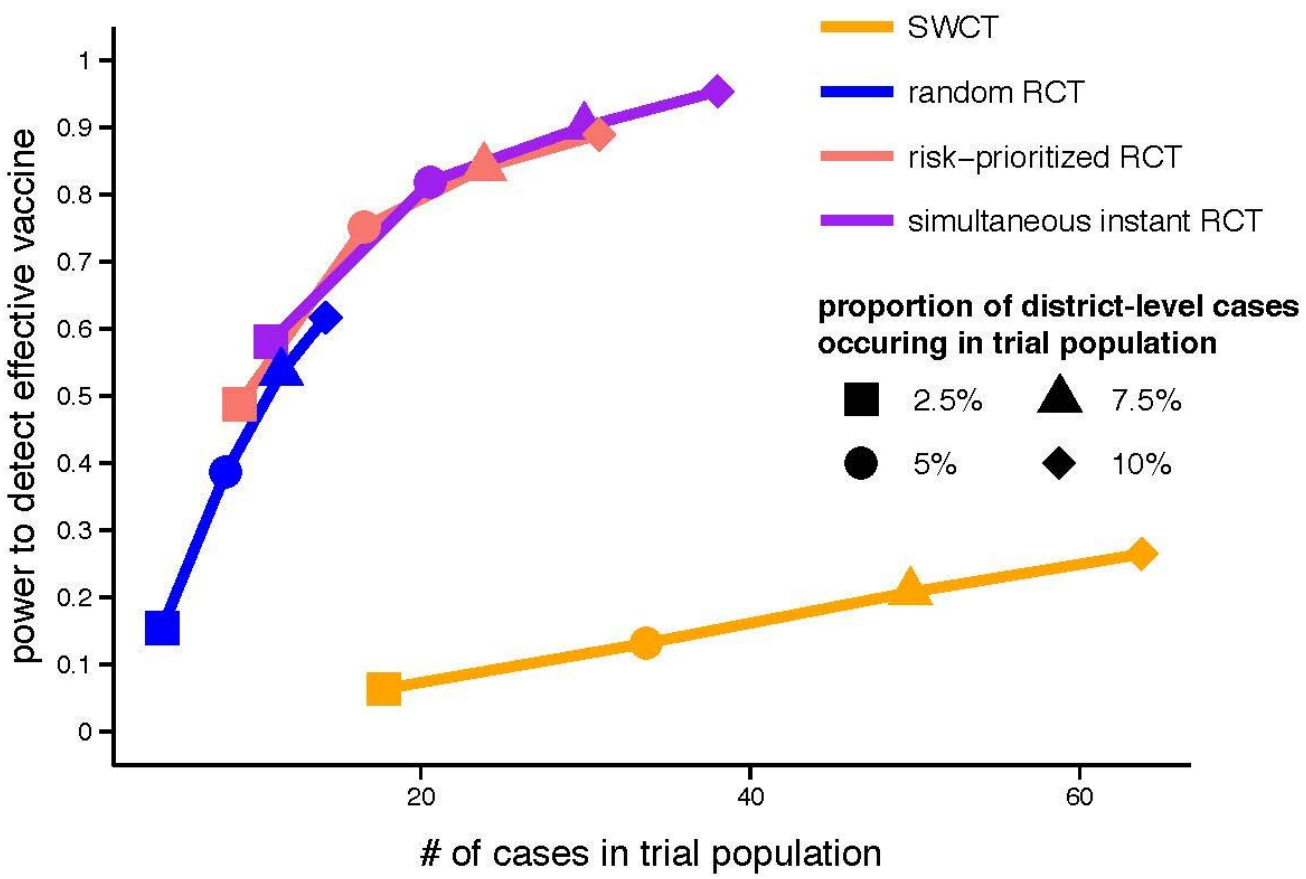

Figure S8. Power by analyzable cases in the trial. Power for a vaccine efficacy of 0.9 shown by the average number of cases in a trial (across 2000 simulations), where each line from left to right represents an assumed 2.5%, 5%, 7.5%, or 10% of district-level infections occurring in the trial population (shapes). RCT and SWCT power are calculated with CoxPH and permutation tests (over CoxPH estimators), respectively. For the same proportion of cases assumed to occur in the trial population, each design still captures a different subset of person-time and events (black lines in Figures 3, S1-S2) and vaccinates different proportions of the population. All RCT designs have the same efficiency (i.e. power per number of cases) and only differ by the number of cases they capture. The simultaneous instant RCT captures the most cases, followed by risk-prioritized designs, and then random-ordered designs. The FRCT is not shown for clarity, because it lies along the same curve as the other RCTs.

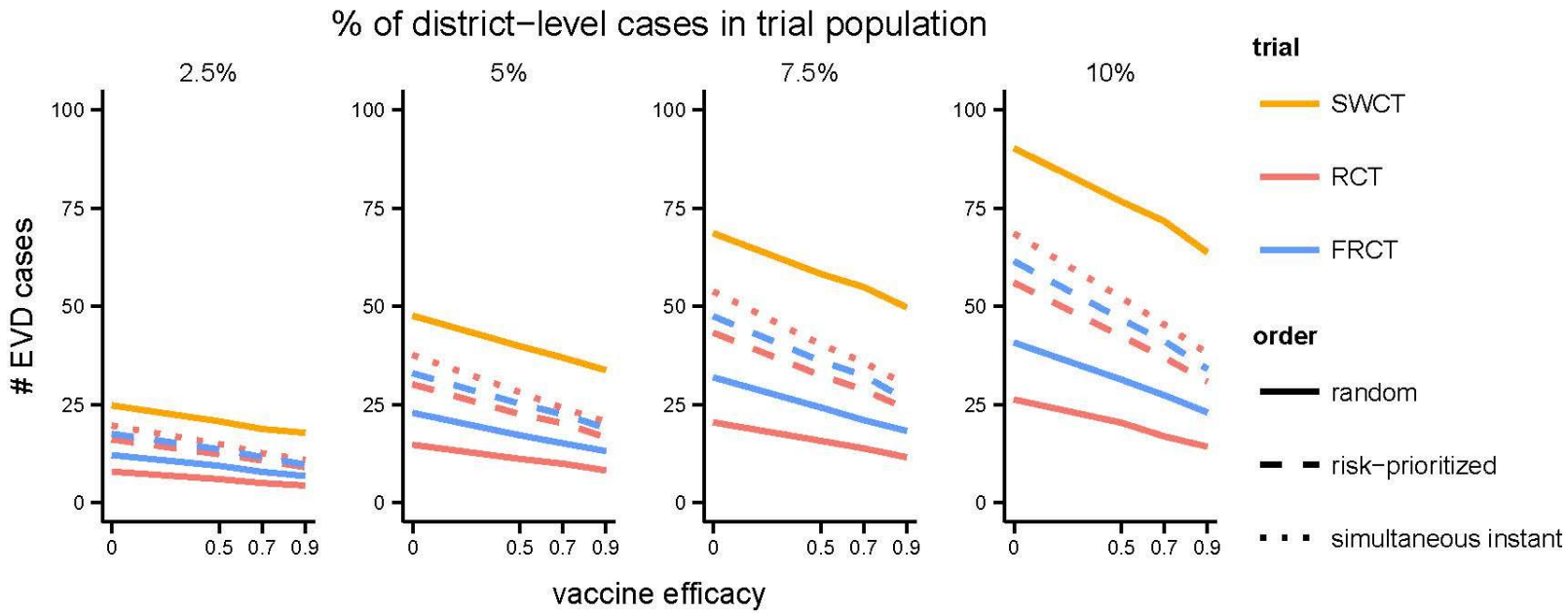

Figure S9. Analyzable cases in trial by trial design. Shown by proportion of district-level cases assumed to be in the trial and the order of cluster vaccination. Cases in clusters of RCTs that have not yet had vaccine rollout are not considered to have occurred within the trial because they cannot be included in a formal analysis. The greater number of cases in the SWCT compared to an RCT simultaneously vaccinating all participants in the vaccine arm at trial initiation arises because the SWCT includes cases before the 21 day vaccine protective delay ends, whereas analyses of an RCT only include cases in clusters in which vaccinated individuals have already been considered to develop protective immunity. This causes a substantial difference in the number of cases due to the rapid incidence decline. However, this does not give much added advantage to the SWCT's power (Figure 4, S8) due to the temporal fluctuations within each cluster, which weaken inference based on before and after comparisons, and the weaker power of cluster- versus individual-level randomization. Increased vaccine efficacy reduces the number of cases in a trial via vaccine-induced prevention of infection.

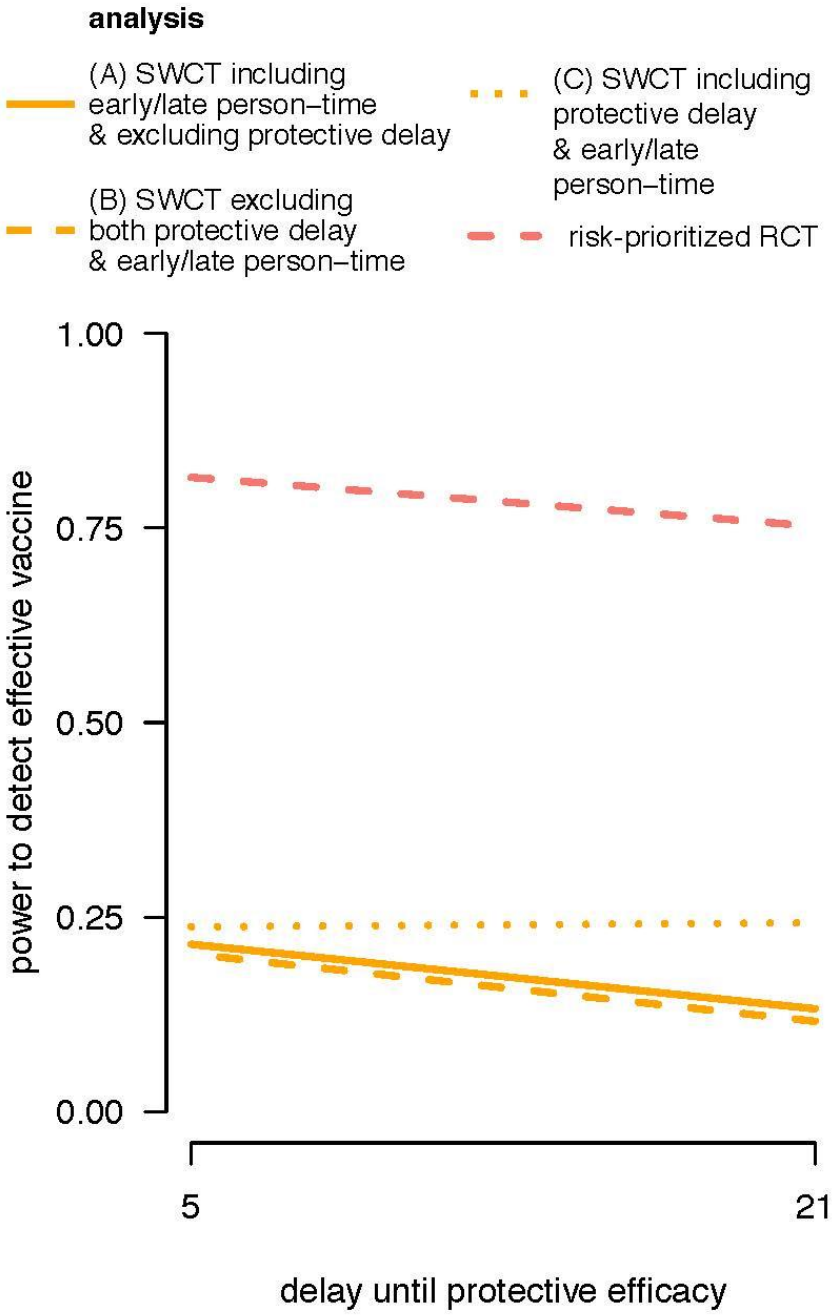

Figure S10. Power by protective delay. Power shown for SWCT and risk-prioritized RCT for vaccines assuming 5-day or 21-day lags from vaccination until protective immunity. Analyses are done under the assumption that this delay is known, and the differences in power shown here primarily result from differences in the amount of vaccinated person-time observed. Power is calculated using CoxPH for the RCT and a permutation test (over CoxPH the estimator) for the SWCT. This can be seen by the relative insensitivity of the SWCT design that includes the protective delay in the analysis to the extra 16 days of seroconversion delay. In this analysis the delay period is included in the analysis as unprotected person-time, and therefore increasing its duration does not reduce person-time analyzed. First, while in an RCT these 16 days of person-time are excluded from the analysis, in an SWCT we include them as unprotected person-time.

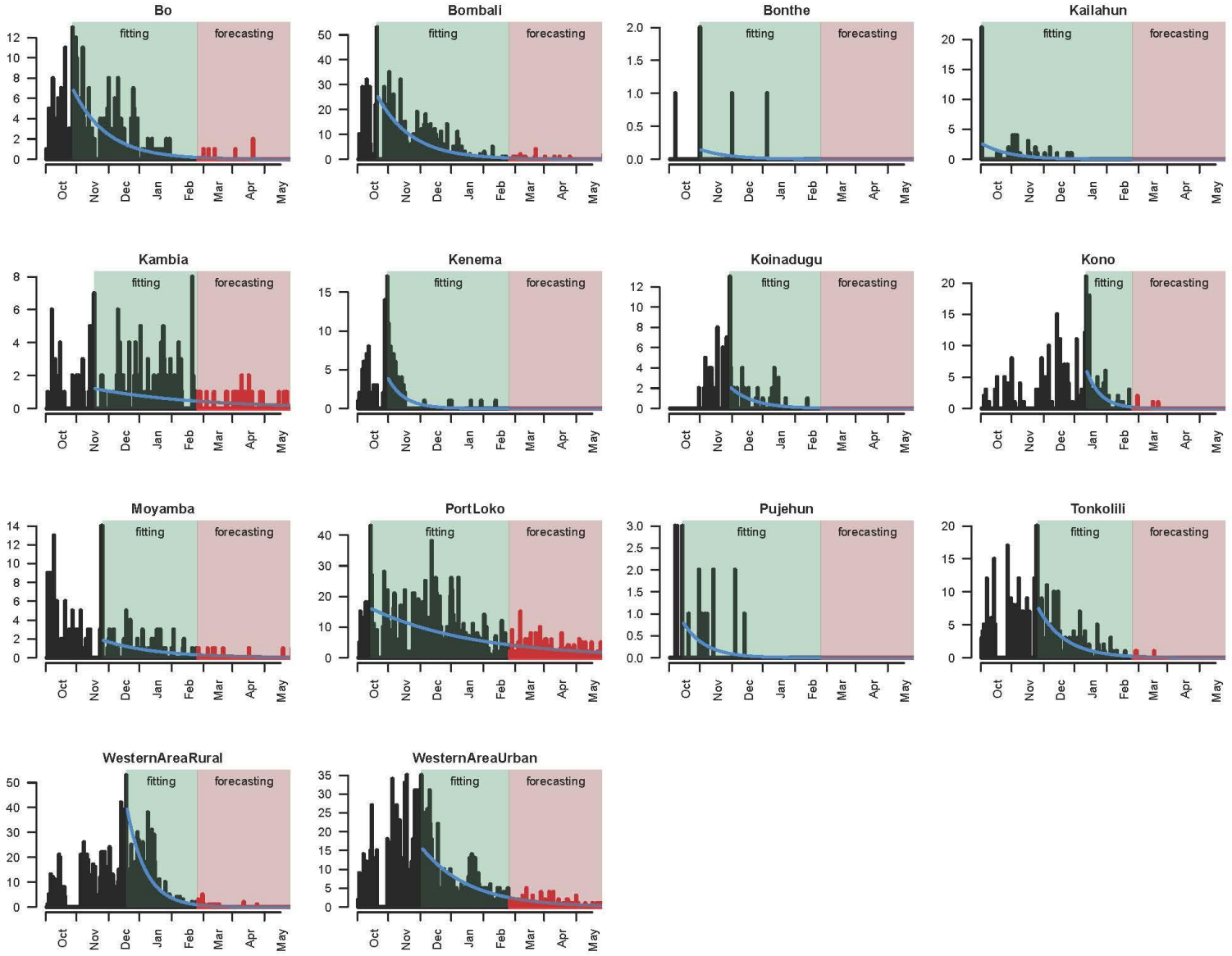

Figure S11. Fitted incidence projection models for all districts. Same as Figure 1 but showing exponential model fits and example forecasts for all districts.