VIMC Model Method

**Underlying spatial incidence rate estimates**

We use the 5 km by 5 km spatial incidence rate estimates from a published study (<https://doi.org/10.1016/S0140-6736(17)33050-7>) for the 35 countries in sub-Saharan Africa. These countries are as follows: Angola, Burundi, Benin, Burkina Faso, Central African Republic, Côte d'Ivoire, Cameroon, Democratic Republic of the Congo, Republic of the Congo, Ethiopia, Ghana, Guinea, Guinea-Bissau, Kenya, Liberia, Madagascar, Mali, Mozambique, Mauritania, Malawi, Namibia, Niger, Nigeria, Rwanda, Senegal, Sierra Leone, Somalia, South Sudan, Chad, Togo, Tanzania, Uganda, South Africa, Zambia, and Zimbabwe.

As for 10 countries outside of sub-Saharan Africa, including Afghanistan, Algeria, Haiti, Iran, Iraq, Nepal, Pakistan, Philippines, Thailand, and Yemen, we use and process the country-level annual cholera cases from WHO Weekly Epidemiological Record (WER). We first calculate the mean of the annual cholera cases from the most recent and available 5 years, divide it by the mean population averaged over these 5 years using demographic data provided by VIMC, cholera incidence rate is spatially homogeneous across the country, and then use the singular annualized incidence rate for the whole country.

For Bangladesh and India specifically, we generate the spatial incidence rate layer in a different way. We have an internal dataset with the spatial incidence rate estimates for Bangladesh; the population-weighted singular incidence rate calculated from this dataset is then used for India given the consideration that India doesn’t keep a good record of cholera cases and India may face a similar cholera infection risk as Bangladesh.

To add stochasticity to this process, we draw random spatial layers from the 1000-layer spatial incidence rate data for the sub-Saharan African countries; in the other countries, we assume the annualized cholera cases follows Poisson distribution and generate random layers accordingly.

**Time-varying incidence rate trend**

Over a period of 101 years (from 2000 to 2100) to simulate in the model, the underlying incidence rate is unlikely to remain constant, we implement the incidence rate trend to factor in the time-dependent variance in the spatial incidence rates.

We assume a deterministic log-linear relation between time (in years) and annual incidence rates using Poisson regression model: . To estimate the model coefficients, a dataset summarizes WHO-reported annual cholera cases for all available countries and years is developed initially. For each individual country among the countries with at least 5 years of WHO annual cholera case data, we use their model coefficients and calculate their annual incidence rates from 2000 to 2100. For countries without adequate WHO data ( 5 years) or with extreme annual incidence rate estimates from the Poisson regression model, we develop another Poisson regression model that aggregates case data from all available countries and gives a global estimate: .

To use the year-dependent incidence rate trend for each country on top of their underlying spatial incidence rate, we assume the yearly incidence rate trend and spatial incidence rates converge in year 2014—for year 2014, the original spatial incidence rate suffices; for any year other than 2014, we calculate a multiplicative multiplier for the yearly incidence rate trend relative to its 2014 estimate and multiply it to the spatial incidence rate as the new spatial estimate for that year. The plot for the year-dependent incidence rate trend multiplier can be found in the supplementary material.

**Outbreak pattern**

The cholera cases are usually reported when a district-level outbreak happens. The static spatial incidence rates along with the incidence rate trend treat all districts the same with respect to outbreak dynamics, disregarding that the random outbreaks that happen in some high-risk districts may drive up the cholera cases and surpass other districts to an extreme degree in some years.

We implement the district-level stochastic cholera outbreak spatial multiplier. We adopt a dataset from a published paper (<https://doi.org/10.1101/2021.10.25.21265347>) that has the outbreak data across different spatial scales in multiple countries from 2010 to 2020, including Angola, Benin, Côte d’Ivoire, Cameroon, Congo – Kinshasa, Congo - Brazzaville Ethiopia, Ghana, Guinea, Guinea-Bissau, Kenya, Mali, Mozambique, Malawi, Namibia, Niger, Nigeria, Sierra Leone, South Sudan, Chad, Tanzania, Uganda, Zambia, Zimbabwe, and Somalia. Different districts across spatial scales with outbreak data are summarized in a table as part of the supplementary material. This dataset has quantitative information about each individual outbreak, including the outbreak attack rate, the name and spatial scale of the district where the outbreak happened, the threshold incidence rate, which is used to determine whether an outbreak happened in one district—an outbreak should have an attack rate well above the threshold incidence rate (more detailed definition of an outbreak period can be found in the paper), the outbreak start time, and the outbreak duration. For countries and districts without outbreak data, we assume that there is no time-varying outbreak pattern. For countries with outbreak data, we only focus on the level-2 administration districts provided they have the most available data, including Benin, Côte d’Ivoire, Cameroon, Congo – Kinshasa, Congo - Brazzaville Ethiopia, Ghana, Guinea, Guinea-Bissau, Kenya, Mali, Malawi, Namibia, Niger, Nigeria, Sierra Leone, South Sudan, Chad, Tanzania, Zambia, Zimbabwe, and Somalia. We first calculate the duration-weighted mean incidence rate over the outbreak periods and non-outbreak periods across all level-2 administration districts, then divide the outbreak and threshold incidence rate by the mean incidence rate to calculate outbreak scalar and threshold scalar for each district. The annual probability of outbreak in each district is calculated by dividing the number of outbreaks observed in that district by the 10-year period of observation.

For each level-2 administration district, the spatial outbreak multipliers are generated over the first decade in the period from 2000 to 2100 and repeatedly used for the rest of the years following the decennial pattern, e.g., the same outbreak multiplier is applied to 2001, 2011, 2021, etc. Given there are 50 sets of stochastic underlying spatial incidence rate, the outbreak multipliers are generated in correspondence to each layer individually. Prior to assigning values, an empty set of spatial multipliers is initialized with the only value “1”. In the process of generating an outbreak multiplier given a certain year, a certain layer of spatial incidence rate, and a certain level-2 administration district, an outbreak check is passed first based on the outbreak probability: if an outbreak is supposed to happen, then the pre-calculated outbreak scalar will be applied to where the district belongs to on the spatial outbreak multiplier while the other districts remain unchanged; if an outbreak is not supposed to happen, the pre-calculated threshold scalar will be applied instead. It’s worth noting that for districts without any temporal outbreak patterns, the initial value “1” will remain on the spatial outbreak multiplier alongside other districts.

**Vaccination campaign strategy, indirect effects, and waning effectiveness**

The vaccination campaign is hypothesized to be undertaken on the district level in each country. First, the level-2 administration districts are ranked based on the size of population affected by cholera. The affected population size is calculated by multiplying the 5 km by 5 km spatial population data in 2020 with the mean spatial incidence rate in each district, which is from the published study (<https://doi.org/10.1016/S0140-6736(17)33050-7>) as well. Second, the vaccines are given to the highest-ranking districts to fully cover their population until none is left. The vaccination coverage data is provided by VIMC, including vaccination year, target population, coverage, and fully vaccinated people (target population times coverage). In each vaccination campaign year, given the limited vaccines, not all level-2 districts can have their whole population fully vaccinated, there are almost always districts not receiving any vaccines and one district partially vaccinated because of the vaccination strategy we implement. Additionally, if the vaccination campaigns are undertaken in adjacent years, we will not vaccinate a vaccinated (fully or partially) district again during the following three years.

The vaccines are assumed to directly reduce the proportion of population susceptible to cholera infection, in the meanwhile, the indirect protection of the vaccines and the waning effectiveness are also considered. To facilitate the mechanism of vaccination campaign bringing indirect protection to the unvaccinated, we assume that a certain vaccination coverage level will reduce the underlying spatial cholera incidence rate in each level-2 administration district. We select two published studies (DOI) that reported a range of cholera vaccine coverage levels and the corresponding incidence rates in each unvaccinated sub-population, combine their data, and develop a logistic regression model based on the data, which will inform us the degree of reduction in the spatial incidence rate given a coverage value from any districts. As for waning effectiveness of the cholera vaccine, we use the data from a published meta-analysis (<https://doi.org/10.1016/S1473-3099(17)30359-6>) that reported both effectiveness and efficacy. First, we subset the data with no missing value for mean efficacy, upper bond of efficacy, lower bond of efficacy, standard error, and the period over which the efficacy was measured after people were fully vaccinated. Second, a simple linear regression is developed with middle time point of the measurement period as the explanatory variable, mean efficacy as the response variable, and the inverse of the standard error squared (variance) as the weight. This model is then used to inform the remaining proportion of protected population a certain number of years after the vaccination campaign. It’s worth noting that we assume the protection from vaccines doesn’t last for more than 5 years.

**Simulation, different settings, and model output**

The simulation is done on the yearly basis from 2000 to 2100 across two counterfactual scenarios: vaccination campaign scenario and no vaccination scenario, each with 50 stochastic runs. For each year, different layers of spatial data and scalars are aligned together and multiplied to estimate the number of cholera cases, including spatial population (integrated with the temporal change from the VIMC population data), spatial default cholera incidence rate, spatial susceptible population proportion (derived from vaccine coverage data), spatial relative reduction in incidence rate due to the indirect protection from vaccination campaign, spatial outbreak-related incidence rate multiplier, and the incidence rate temporal trend scalar. The stochastic simulations are facilitated multi-layer spatial data, including the spatial incidence rate and the spatial outbreak-related incidence rate multiplier. Each layer is used for one stochastic run. Under the no vaccination scenario and among the years when the vaccines have no protective effects on the population under the vaccination campaign scenario, the overall population is assumed to be susceptible to cholera infection.

We consider four different settings, each of which has two scenarios and 100 stochastic simulations. These four settings are different in terms of whether the outbreak-related incidence rate multiplier and the incidence rate temporal trend scalar are applied. Additionally, they also have different underlying spatial incidence rate given 50 new layers of spatial incidence rate data are randomly drawn from the source for each setting. On the side note, the same set of outbreak-related incidence rate multipliers are used for the two settings that have the outbreak-related incidence rate multiplier applied for the purpose of efficiency.

With the estimated number of cholera cases in each country, we can then calculate the number of death and disability-adjusted life years (DALYs) related to cholera infection. The country-specific case fatality ratio (CFR) data is from WHO (source), which is used to multiply with the number of cases to calculate the number of deaths. For the countries without CFR data or with CFR data higher than 0.07, we aggregate all the cholera case and death data from WHO and calculate an overall CFR, which is then used on those countries. Cholera-related DALYs is calculated using the data provided by VIMC (source).