

Sampling in Difficult Settings: A Simulation Study Comparing Several Sampling Methods

Harry S. Shannon (✉ shannonh@mcmaster.ca)

McMaster University Faculty of Health Sciences <https://orcid.org/0000-0002-7507-427X>

Patrick D. Emond

McMaster University

Benjamin M. Bolker

McMaster University

Román Viveros-Aguilera

McMaster University

Research article

Keywords: sampling methods, Extended Program on Immunization (EPI), virtual populations, computer simulation, Global Positioning Systems (GPS)

Posted Date: May 11th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-85299/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background:

Taking a representative sample to determine prevalence of variables like disease is difficult when little is known about the target population. Several methods have been proposed, including a recent revision of the World Health Organization's Extended Program on Immunization (EPI) surveys. The original EPI method samples towns as Primary Sampling Units (PSUs) with probability proportional to size and uses a nearest neighbour approach to sample households within PSUs. The new version samples from smaller PSUs and conducts a probability sample of households within those PSUs. Other techniques use satellite images and Global Positioning Systems to sample within towns from circles around randomly identified points ('Circle' method) or from randomly sampled squares in a superimposed grid ('Square' method). We compared these sampling methods in multiple virtual populations using computer simulation.

Methods:

We constructed 50 virtual populations with varying characteristics. Populations comprised about a million people across 300 towns. We created three populations with different prevalences of disease but with uniform characteristics across each population. We created a binary exposure variable and allocated disease statuses to individuals assuming different relative risks (RRs) of exposure. We simulated thirteen methods of sampling: simple random sampling; the original EPI method and variants; the Square and Circle methods; and the new EPI method.

For each population, each sampling method, and each of three sample sizes per PSU (7, 15, and 30), we simulated 1,000 samples. For most sampling methods, the PSUs were towns. We conducted simulations using the same 30 PSUs and using a freshly-chosen set of PSUs. For each simulation we estimated prevalence and RRs and combined the bias and variance of the 1,000 samples to compute the Root Mean Squared Error (RMSE).

Results:

The Circle and Square methods produced almost identical results, so we report only the Square method results. Apart from simple random sampling, the RMSE for the Square method was almost universally best for estimating prevalence, and generally best when estimating relative risks. The revised EPI approach was worse, but generally better than the original EPI.

Conclusions:

The Square method is recommended as statistically optimal, unless practical considerations favour another approach.

Background

Surveys across the world are conducted for many reasons, such as quantifying population prevalence of disease or immunization rates. Conducting these surveys can be challenging when background information on the population of interest is limited.

Various survey methods have been proposed for such low-information scenarios, some of which have been applied in the field. The World Health Organization (WHO) developed a methodology to estimate immunization rates in young children, the Extended Program on Immunization (EPI) [1]. The EPI method uses a two-stage methodology, in which towns are the Primary Sampling Units (PSUs) selected (with probability proportional to PSU size) at the first stage, while individual households are chosen within each PSU via a nearest-neighbour strategy. Others have used technology that has since become available, including satellite images and Global Positioning Systems (GPS) [2] or satellite images and GIS sampling [3].

The original EPI approach was discussed in several papers. Some authors have noted the EPI's limitations, including the inability to estimate sampling probabilities, and suggested modifications to overcome those limitations [e.g., 4–6]. In 2018, the WHO published a revised EPI approach, including substantial changes to the methodology [7]. The new technique identifies Enumeration Areas (EAs), samples EAs as the PSUs, and samples all or a random sample of eligible people within the chosen EAs. We describe these approaches in more detail in the Methods section. We comment on the relationship with various methods that use 'gridded' population sampling in the Discussion.

Some authors conducted simulations to determine if the original EPI method was 'good enough,' i.e., whether the bias and variance of the estimates were sufficiently small for the survey's purposes [e.g., 8, 9]. However, to our knowledge, the more recently developed methods have not been tested in this way and only one study we are aware of has gone beyond prevalence and examined relative risks (RRs) [10].

We have conducted a simulation study of various approaches including both the older and the newer EPI methods. Our aims were (i) to conduct simulation studies to determine the bias and precision of different sampling methods and (ii) to determine the circumstances in which one method may be preferred over others. We have conducted an extensive study, creating multiple large populations with varying characteristics and comparing many methods.

As noted above, the binary outcome of interest can be presence or absence of a disease or some other characteristic such as vaccination status. Henceforth, we refer to the outcome as 'Disease.'

Methods

The broad approach was as follows:

- Create many virtual populations with known characteristics (parameters), including allocation of disease or vaccination status, an exposure and disease status for different relative risks (RRs) from that exposure.

- Simulate different sampling methods to take multiple samples from the populations
- Estimate the prevalence of disease and the RRs from an exposure for each sample
- For each method, combine the bias and variance of the samples to compute the Root Mean Square Error (RMSE) of the estimates
- Compare the RMSEs for the different sampling methods both overall and in relation to the population characteristics

Creating the virtual populations

The simulation program was written to be extremely flexible. A variety of parameters was chosen, as we attempted to mimic how those parameters might vary in real life. We varied parameters for the overall populations and for characteristics of towns, households, and individuals within populations. To consider a broad range of many different parameters we used a 'Latin hypercube' approach [11], treating the parameters as measures that varied in small increments within a pre-specified range and ensuring unique combinations of the parameters. The technique is in effect a stochastic form of fractional factorial design that works well with large numbers of parameters.

For each simulated population, we randomly sampled one of the possible values for each parameter without replacement. For example, for the mean sizes of households the range was from 2 to 5, varying by units of 0.06. Since we created 50 populations with characteristics varying between and within the towns, we allowed 50 values for the parameters, and the Latin Hypercube approach ensured we used each of those values in exactly one simulated population. Other population parameters included the target disease prevalence, number of disease pockets per town (see below), and prevalence of exposure. The values used were chosen as for household size.

Each population created was distributed among cities, towns and villages (henceforth, simply 'towns') using a Pareto distribution. We created 300 towns, with population sizes between 400 and 250,000. Each town was geographically represented as a square, 10,000 by 10,000 units. Given a parameter value for a population, the actual value for a particular town was randomly chosen from a normal distribution centred at the population value with a small variance. Within each town, we divided the area into 100 smaller squares (a 10 x 10 grid) each 1,000 by 1,000 units, labelling the axes x and y. The values of x and y were used to determine the overall characteristics of people living in the sub-area. The first determination was the range in the density between the most and the least densely populated sub-areas. The density varied linearly with each of x and y, so that the minimum and maximum densities were at opposite corners of each town.

The households were placed within each square by randomly choosing spatial coordinates within the appropriate range. We did not require a minimum distance between households; any two households close together could be considered to be part of a multi-residence building. The number of people in a household was randomly determined, based on the hypercube value for the mean number per household, using a zero-truncated Poisson distribution. The first two people in the household were taken to be adults,

and additional members were designated as children. Using the linear function that determined the population density, households were added until the sub-area had the predetermined number of people. Other characteristics, such as household income or age (adult vs. child), were determined using the other parameters. Additional file 1 shows the parameters used in the simulations, and the ranges of possible values allowed.

We incorporated 'pocketing', the presence of small areas with particularly high prevalence, representing a local spread of infection. This was done by randomly identifying points that were the centres of pockets. The number of pockets per town was randomly chosen for each population. Each pocket added to the risk of disease for everyone in the population. The risk declined rapidly with distance from the centre of the pocket, so for most people the additional risk was minimal. The program allowed for different numbers of pockets and for differing functions and rates of decline with distance of the additional risk.

The allocation of disease status to each individual was based on the risk, which was in turn based on several factors. Once the background disease risk for a sub-area was determined, several additional factors played a role, including age and household income. Each person's disease status was determined randomly based on the calculated binomial probability, p . The disease determination algorithm in Additional file 2 provides further detail. In running the program, the prevalence of disease in a population was not equal to the target value that had been chosen.

We also incorporated bivariate relationships between variables representing an exposure and a disease. The likelihood of exposure varied across the population depending on the location of the household. We considered relative risks (RRs) of 1.0, 1.5, 2.0, and 3.0. To program these, we assigned a different disease for each RR; for Disease 1 we had $RR = 1.0$, for Disease 2 we had $RR = 1.5$, etc.

Each disease status for individuals was based on the exposure level (present / absent), the background disease risk, and the relative risk. For example, if the background disease risk was 0.1, and the the relative risk of Disease 3 was 2.0, the risk was the product, 0.2. Individuals were assigned Disease 3 status randomly, with binomial probability of 0.2. When the background prevalence and the RR were high, the product could produce a probability greater than 1, so we 'capped' probabilities at 0.9. As with prevalence, the actual RRs differed from the target values.

Three additional populations were created with different prevalences but no variation in the parameters across or within the towns.

Applying the sampling methods

The methods used a cluster sampling design. In all methods except one (labelled 'NewEPI'), the PSUs were towns. PSUs are selected using Probability Proportional to Size (PPS). In practice, this is Probability Proportional to Estimated Size (PPES) since the PSU sizes are not known exactly. Households are selected in each PSU until the specified sample size of individuals is reached. The methods differ in how they identify the sample within the PSU, and the ones we programmed had been used or proposed in the literature. Towns can vary considerably in size, so when they form the PSUs, the large ones typically are

chosen with probability 1, i.e., with certainty. Sometimes the PPS selection method can choose a town more than once. If the town is chosen k times, then k samples are taken from the town.

The sampling methods within the PSUs were:

Simple random sampling – ‘Random’

Simple random sampling (SRS) selects households with equal probability within PSUs. While logistically impractical in real-life populations, SRS was our standard for comparisons of the methods.

The original EPI method – ‘OldEPI’

The original Extended Program on Immunization (EPI) approach [1] identified a landmark in the centre of a town, and chose a random direction from that landmark. Interviewers walked along that line to the edge of the town, identifying buildings on the line. One building was randomly chosen as the starting point. The household in that building (if any) was asked to participate in the survey. The next household chosen was the next closest household residence. This ‘nearest neighbour’ identification continued until the required sample size was reached. EPI aimed to estimate the proportion of young children who had been immunized against some disease(s), so limited interviews to households with eligible children.

In practice buildings occupy an area in two dimensions, whereas we placed each building at a point. So instead of drawing a line from the centre of the town to the edge, we drew a strip 100 units wide, symmetrical about the random direction, and identified buildings in that strip. As well, we determined an arc of 60° , symmetrical about the random direction. One building in the strip or arc was randomly selected as the starting point. We labelled these approaches ‘Strip’ and ‘Arc’.

The original EPI method, choosing every 3rd or 5th household – ‘EPI3’, ‘EPI5’

By choosing every 3rd or 5th household, it is expected that the sample will come from a wider geographical area, and thereby be less susceptible to biases [4]. We operationalized this approach by using the OldEPI method of successively identifying the nearest neighbour, but only sampling every 3rd or 5th household. As with OldEPI, we identified the first household using both Strip and Arc techniques.

Half sample from centre, half from periphery – ‘Peri’

This method [4] also aims to obtain the sample from a wider geographic area than OldEPI. A random direction is chosen from the centre of a town, and the first household in that direction is sampled. The nearest neighbour method then obtains half the target sample for that town. A new random direction is found and the last household along that line to the edge of the town is sampled. The remainder of the sample is found using the nearest neighbour approach. We used strips rather than lines as described above to identify the two ‘starting’ households.

Selecting parts of the sample from each quadrant – ‘Quad’

A further method designed to provide geographic dispersion of the sample divides the town into four quadrants, and applies the OldEPI method to each of them, replacing the central landmarks with the

centres of the quadrants [4]. A quarter of the sample is taken from each quadrant.

The Peri and Quad simulations did not yield equal numbers in the different areas when the target sample size in any town was not divisible by two (Peri) or four (Quad). We ensured the split was as even as possible, randomly determining which areas would have an extra 'participant'.

Using a grid to identify the initial household - 'Grid'

One approach that avoids the use of a central landmark was to use a rectangular grid 'superimposed onto a scanned streetmap of the quartier,' randomly select an intersection of the grid lines, and identify the 'closest compound to the right' as the starting point for the EPI method [12].

Circles about random points - 'Circle'

Random GPS points are chosen in each town, a circle of fixed radius is drawn on an aerial image around each point, the buildings within the circle are identified from the image, one building is randomly chosen, and a household in the building is interviewed. This continues until the target sample size is reached [2]. We applied this technique, using circles of radius 100 units.

Square grid - 'Square'

The potential for overlapping circles in the above procedure potentially complicates the estimation of probabilities of selection. Instead, a grid of squares can be imposed on the image of the town to avoid overlap. (In practice, adjustment to town boundaries is necessary.) Grid squares can be randomly chosen, and a household randomly selected from those squares [13]. In this study, we used a 64 x 64 grid of squares over each town. In practice, buildings can overlap the edges of the squares. Since we defined buildings as points, we did not have this problem, although buildings could land on an edge dividing squares. We specified that a building on the south or west edge of a square was deemed to be in that square.

The new EPI method - 'NewEPI'

The revised version of EPI has developed a different method of conducting vaccination coverage surveys than OldEPI [7]. In brief, the approach is as follows: The population under study is divided into Enumeration Areas (or equivalent), which are of fairly similar size. Large EAs are segmented into smaller areas, while small EAs are combined with contiguous areas to ensure that the EAs differ in size by no more than twofold.

EAs are then selected using PPES. Once the sample of EAs has been chosen, the households in each are enumerated. A random sample within each EA is chosen, up to the required sample size.

The method differs from the other approaches, which all use towns as the clusters. By enumerating all households in the EAs, the probability of selection can be estimated, unlike in OldEPI. In our virtual populations, we did not create EAs. Instead, we constructed EAs by dividing towns into rectangular areas with between 50 and 100 households.

Other researchers have proposed alternatives to EAs, since the most recent censuses in lower-income countries may be very out-of-date. Several 'gridded' population datasets have been constructed. They use a statistical model with available spatial data to estimate populations in small grid cells. Thomson et al. describe 43 surveys that have used gridded datasets [14]. Since these studies identify PSUs analogous to EAs, these gridded population surveys can be considered equivalent to the NewEPI methodology for our purposes.

Yet another approach is to 'segment' clusters [5, 6] (sometimes done in surveys using gridded data). The segmentation process is done by teams in the field. The teams produce rough maps of the clusters and divide them into smaller areas (segments) of roughly equal size. One segment is randomly chosen, and a complete list of households obtained, from which the required sample is randomly chosen. However, this procedure requires relatively small clusters, whose size is known from a recent census. Thus, in Turner et al.'s example, the clusters were mostly 250–300 households [6]. The NewEPI technique uses EAs which likely match the sort of clusters needed for this approach, so we have in effect included sampling using segments.

In summary, we included 13 sampling methods: Random, OldEPI strip, OldEPI arc, EPI3 strip, EPI3 arc, EPI5 strip, EPI5 arc, Peri, Quad, Grid, Circle, Square, NewEPI. Since the strip and arc procedures to identify the first household in a town yielded almost identical results, we show results only for the former. The results for the Square and Circle methods were also very similar; we present results only for the former.

Sample size

The OldEPI methodology initially used a sample size of 30 towns and 7 children per town (30 x 7). It was estimated that this would provide 95% confidence intervals of $\pm 10\%$ for a prevalence of 0.5, considered sufficiently narrow for the purposes of the surveys [7]. Other studies have used 30 clusters and 30 people per cluster (30 x 30). We also included samples of 15 per cluster to provide an intermediate value. Thus we simulated samples of 30 x 7 (210), 30 x 15 (450), and 30 x 30 (900). Within any PSU, samples of 7, 15, and 30 children were taken independently.

Simulated sampling

The sampling methods we included were programmed into our simulations. After a population was created, we used two ways to identify PSUs for the sample. The first ('same PSUs') identified 30 PSUs which were used to obtain all 1,000 sets of simulated samples. The second approach ('resampling') took a fresh sample of PSUs each time a new set of samples was taken. A set consisted of the three samples sizes (210, 450, and 900) x nine sampling methods, i.e., 27 samples. One thousand sets of samples were taken.

Analysis

Calculating probabilities of selection:

The OldEPI method and its variants assume that the estimated sizes of towns are correct, or at least proportional to the actual sizes. If so, the probability of choosing any person in the population is

constant, and no sampling weights are needed. OldEPI also assumes that within the town a simple random sample is taken, which is not the case, so it is further assumed that any flaws in the sampling will even out over 30 towns. We therefore did not weight our simulated samples that applied the OldEPI methodology or any of its variants. Likewise, the NewEPI conducts PP(E)S and takes a fixed number of people per EA, so the samples are self-weighted.

For the Circle and Square sampling, we estimated the sampling probabilities. We first computed the probability that the town was chosen, p_t . Since the town had been selected using PPS, we took this as $n_i \cdot k/N$, where n_i = size of town i, k = number of towns (i.e., PSUs) selected (30), and N = total population.¹

For the Circle approach, we calculated the fraction of the total area of town i that the circles selected took up. This was calculated as $p_a = (100)^2 \cdot \pi \cdot c_i / (10000)^2$, where c_i = # circles selected, and the radius of each circle was 100 units, while the town size was 10,000 x 10,000 units. The number of circles could differ by town since the calculation includes the empty circles. We assumed that the circles did not overlap.

This gave the probability of selecting the circles, given the town. We included in the numerator the areas of circles that did not contain any households and circles for which the chosen household did not include an eligible person. The inverse of n_{ih} , the total number of households in each circle in town i, gave the probability of choosing the household given the circle, $p_{ih} = 1/n_{ih}$. The equivalent, somewhat simpler, calculations were made for the Square sampling.

Since we sampled all eligible people in the household, the probability of selecting an individual given the household was one. The product of the probabilities was the sampling probability for an individual in the sample, $p_{\text{indiv}} = p_t \cdot p_a \cdot p_{ih}$. The sampling weight was the inverse of the individual sampling probability, $w_{\text{indiv}} = 1/p_{\text{indiv}}$.

Calculating Prevalences and Relative Risks

For any particular sample, the prevalence of disease was calculated, as were the RRs for the four diseases noted above. Sampling weights were applied when appropriate. The variances of the 1,000 values of the prevalences and RRs were computed for each cluster size and sampling method.

Since the population was known, the true prevalences and RRs were known. Thus we also computed the error of any sample (sample estimate – true population value), and took the mean of those values to compute the bias. We then computed the variance of the means across the 1,000 simulated samples. We used the bias and variance to compute the Mean Squared Error (MSE), where

$$MSE = (Bias)^2 + Variance = Mean\{(Estimate - TruePopulationValue)^2\}$$

The Root Mean Squared Error (RMSE), the square root of the MSE, was the measure for comparing the sampling methods.

Overall Comparisons of the Sampling Methods

We made the comparisons in two ways:

First, for each population and sample size we ranked the RMSEs. The lowest RMSE was given the rank 1. The next lowest was given the rank 2, etc. We calculated the mean rank for each sampling method and sample size across the 53 populations. For each population and sample size we also took the ratio of the RMSE for the sampling method to the RMSE for Simple Random Sampling. We then took the mean of these ratios for the 53 populations, and compared the means across the sampling methods.

Second, for each population and sample size we took the ratio of the RMSE for the sampling method to the RMSE for Simple Random Sampling. We then took the mean of these ratios for the 53 populations, and compared the means across the sampling methods.

Impact of the Population Parameters

We also wanted to learn how the RMSE ranking varied with different values of the parameters used to construct the populations. We constructed graphs showing the RMSEs for the different methods in relation to the parameter values. We smoothed the plots using generalized additive models.

Computer processing

Before conducting final runs, many test runs were carried out and results were checked to ensure they were consistent with theory and internally. Bugs in the program were identified and fixed.

The creation of the populations and simulations of sampling were conducted on a modern high-performance cluster - we used SHARCNET, a high-performance computational resource supported by a consortium of Ontario universities [15]. The two final runs took approximately 380 processor hours.

Results

As noted above, since the results were very similar for the strip and arc procedures for finding the first household we show data only for 'strip.' Similarly, since the Square and Circle methods gave nearly identical results, we only include those for 'Square.' Further, for clarity, we limit the number of methods compared in the graphs.

Mean Ranks

The tables of mean ranks are based on the ranking of each sampling method's RMSEs for each of the 53 populations. Tables 1 and 2 show the mean ranks for when the relative risks are 1.0 and 3.0, and for when the same PSUs were used for each of the 1,000 simulations. The data for RR = 1.5 or 2.0 and for when PSUs were resampled for each simulation are in Additional file 3. They show very similar patterns.

Table 1
Mean ranks of RMSEs for relative risk = 1.0 and same PSUs are sampled

Sampling method	Mean ranks when estimating					
	Prevalence			Relative Risk (RR)		
	<i>n = 7</i>	<i>15</i>	<i>30</i>	<i>n = 7</i>	<i>15</i>	<i>30</i>
NewEPI	4.2	4.2	4.6	5.1	5.5	5.5
Grid	5.8	6.1	6.0	6.4	6.8	7.2
Peri	3.7	4.0	3.9	3.3	3.9	3.9
Quad	3.4	3.2	3.6	2.3	2.2	2.2
Square	1.5	1.2	1.3	3.5	2.3	1.8
OldEPI	6.7	7.1	7.0	5.5	5.7	6.2
EPI3	5.9	5.7	5.4	4.8	4.8	4.8
EPI5	4.8	4.5	4.2	5.1	4.8	4.3

Note: RMSE = Root Mean Squared Error. PSU = Primary Sampling Unit. For this and other tables of rankings, a low ranking represents a lower RMSE, so is 'better'. (1 = lowest RMSE ratio, 8 = highest RMSE ratio.) The first three columns show the mean rankings for RMSEs of prevalence estimates for the three sample sizes within clusters ($n = 7, 15, \text{ or } 30$). The other three columns show the mean rankings for the RMSE ratios of estimates of relative risks. See text for description of sampling methods and Additional file 3 for full set of tables.

For estimates of prevalence, the Square method is clearly the best, with rankings lower than (better than) those of other methods. The NewEPI and two variants of the OldEPI – Peri and Quad – are similar. The OldEPI methods are generally worse, especially when they sample the next nearest household; when they choose the 5th nearest household, the rankings are better. Overall, the rankings do not change much with sample size.

Table 2
Mean ranks of RMSEs for relative risk = 3.0 and same PSUs are sampled

Sampling method	Mean ranks when estimating					
	Prevalence (RR)			Relative Risk		
	<i>n</i> = 7	15	30	<i>n</i> = 7	15	30
NewEPI	4.2	4.4	4.3	5.5	5.9	5.5
Grid	5.4	5.7	5.8	6.5	6.7	6.9
Peri	3.4	3.4	3.3	3.1	3.5	3.7
Quad	3.7	3.5	4.0	2.6	1.9	2.7
Square	1.7	1.5	1.6	3.2	2.4	1.7
OldEPI	6.7	6.9	7.1	5.7	5.8	6.4
EPI3	5.7	5.7	5.5	4.8	5.1	5.0
EPI5	5.1	4.8	4.5	4.6	4.7	4.0
See footnote to Table 1.						

For estimates of relative risk, the picture is a little different. For the smallest sample size (7 per PSU) the Quad method, which guarantees that some of each town's sample will come from each quadrant of the town, performs well. For the larger sample sizes (15 or 30 per cluster) the Square technique is roughly on a par with the Quad approach.

The Old EPI methods, even when taking the third or fifth nearest house, have poor rankings. The NewEPI method is generally better than the older EPI methods, especially for estimating prevalence.

Means of Ratios of RMSEs

Simple Random Sampling (SRS) is our basis for comparison of the sampling techniques, so we computed the ratios of the RMSEs to the RMSE for SRS. The means of these ratios are shown in Tables 3 and 4 for RR = 1.0 and RR = 3.0 when the same PSUs were used for the 1,000 simulations, and graphs of the results are shown in Fig. 1.

Table 3
Mean ratios of RMSEs for relative risk = 1.0 and same PSUs are sampled

Sampling method	Mean ratios when estimating					
	Prevalence Risk (RR)			Relative		
	<i>n = 7</i>	<i>15</i>	<i>30</i>	<i>n = 7</i>	<i>15</i>	<i>30</i>
NewEPI	1.18	1.41	1.62	1.08	1.31	1.44
Grid	1.28	1.55	1.92	1.35	1.46	1.68
Peri	1.19	1.43	1.74	1.25	1.20	1.35
Quad	1.23	1.43	1.75	1.33	1.05	1.16
Square	1.01	1.04	1.07	1.22	1.00	0.99
OldEPI	1.39	1.73	2.15	1.49	1.34	1.55
EPI3	1.37	1.64	1.97	1.13	1.28	1.44
EPI5	1.33	1.57	1.85	1.45	1.27	1.38

Note: RMSE = Root Mean Squared Error. PSU = Primary Sampling Unit. The first three columns of data show the mean RMSE ratios (ratio of RMSE for the sampling method: RMSE for simple random sampling) for the prevalence estimates for the three sample sizes within clusters ($n = 7, 15, \text{ or } 30$). The last three columns show the mean RMSE ratios for estimates of relative risks. The same towns were sampled for all 1,000 simulations.

The Square method is always better – it has the lowest mean ratios, close to 1, indicating that the RMSEs are similar to those from SRS. Notably, for prevalence, all the other methods have mean ratios that increase with sample size per PSU. The increase in the ratios indicates that the EPI methods benefit less from larger sample sizes, likely because of intra-cluster correlation due to the homogeneity of people in neighbourhoods. This is not surprising for the OldEPI methods, since they take close(st) neighbours when sampling within PSUs. The ratios were always smaller for a bigger gap between households selected (nearest neighbour vs. 3rd nearest vs. 5th nearest), consistent with the interpretation that closer households are more similar to each other than to more distant households.

Table 4
Mean ratios of RMSEs when the relative risk is 3.0 and the PSUs sampled are the same for all simulations

Sampling method	Mean ratios when estimating					
	Prevalence			Relative Risk (RR)		
	<i>n</i> = 7	15	30	<i>n</i> = 7	15	30
NewEPI	1.26	1.55	1.83	1.15	1.32	1.41
Grid	1.35	1.69	2.12	1.23	1.45	1.67
Peri	1.23	1.48	1.81	1.00	1.13	1.32
Quad	1.35	1.63	2.01	1.00	1.03	1.20
Square	1.03	1.06	1.11	1.00	0.98	0.99
OldEPI	1.52	1.94	2.43	1.16	1.32	1.58
EPI3	1.49	1.84	2.24	1.12	1.27	1.45
EPI5	1.46	1.79	2.14	1.11	1.26	1.36
See footnote to Table 3.						

One might have expected the Quad and Peri approaches to be relatively free of this property, since they sample from different areas of the towns, but they were not. Since the NewEPI takes random samples, it might have avoided the problem, but it did not – possibly because the Enumeration Areas are small and homogeneous, so larger samples within the EAs will still choose people very similar to those in smaller samples.

For estimates of relative risk, the Square method performs very well; the mean MSE ratios are mostly close to 1.0, for all three sample sizes. The Peri and Quad procedures are sometimes - but not always - comparable in having low mean ratios. The patterns for all the EPI methods are not consistent when RR = 1.0 - sometimes the mean ratios decrease with larger sample size, sometimes they increase. When RR = 3.0, the mean ratios for EPI methods always increase with larger sample sizes.

In summary, the Square technique is the best, demonstrated by the low mean ranks of the RMSEs and the low mean RMSE ratios. Moreover, the mean RMSE ratios are almost always close to 1, showing it performs similarly to SRS.

Impact of parameter values

The RMSEs for the Square method were ranked the best for the great majority of the 50 populations when prevalence was estimated – typically for more than 40, or even 45, of the populations. For RR estimation, this was generally the case, but not as consistently for *n* = 7 per PSU, when the Quad or Peri approach was quite often the best. Thus when the interest is in prevalence, we did not expect that examining the

relationship between the RMSEs and parameter values (which determined the populations) would identify circumstances when different methods would be preferable. Still, for completeness, we looked at the relationships.

We explored when each method would be preferable by examining graphs of the mean RMSE ratios in relation to parameter values (an example is shown in Fig. 2). Again, they generally show the lowest (best) values for the Square method, especially for $n = 15$ and $n = 30$ per PSU. In general, the relative performance of the various sampling techniques evaluated stays the same across the entire range of parameter values considered.

Discussion

Summary of main results

Our simulations found that the Square method was the best, as measured by the lower mean ranks of the RMSEs and the lower mean RMSE ratios. Under some circumstances, the Peri and Quad approaches, which sample from two or four areas of the towns, respectively, perform quite well, better than the standard EPI method (confirming previous simulations [4]), but not as well as the GPS techniques. The NewEPI is usually better than OldEPI, especially when estimating prevalence.

The superiority of the Square technique was confirmed by the estimates of RMSEs in relation to population parameters which revealed that there were no particular parameters (i.e., no population type or variability within towns) for which other methods were better.

Commentary

Some of the procedures proposed to overcome the known limitations of the original EPI ('OldEPI') were improvements but have their own limitations. For example, our results suggest that the homogeneity of people within small areas will produce relatively high design effects, and increasing sample size will produce only a modest improvement in precision. This applies also to the NewEPI approach, which divides the whole population into small Enumeration Areas. It would be possible to adapt some approaches in the absence of information on the target population, for example, a recently formed informal refugee camp. For example, drones or other technology can ensure any aerial images needed are up-to-date. This would be even more feasible if software can recognize buildings or tents on the ground.

We did not include simulations in which we varied the number of PSUs in our samples or the sample size per PSU. Given our results, though, we expect that methods that take their samples from small areas within a larger PSU are likely to have significant design effects, especially if they take fewer clusters and relatively large sample sizes per cluster. The Square methodology can lead to more variability in the sample weights, which would increase the variance of the estimates of prevalence and RR, yet the RMSEs were still smaller than for the other sampling methodologies.

One possible disadvantage of the Square method is that for geographically larger towns there may be a good deal of travel required to reach all the sampled households, while EPI samples are in a small geographic area. Still, at times the dispersion across a town may be an advantage if there are concerns about the security of interviewers. With the Square technique, the interviewers can enter and leave areas quickly, rather than spend time finding and interviewing several households in a small neighbourhood.

Strengths and limitations of our work

Our study has several strengths. We attempted to create realistic populations, whose characteristics varied across towns. We included multiple populations, which simulations using real data cannot. We included many sampling methods, including some that have been proposed but to our knowledge not used in practice, and the Square and Circle approaches which have not been included in previous simulations. We added the NewEPI method that has only recently been developed. Finally, to our knowledge only one previous study has simulated relative risks [10].

Of course, our study has limitations.

The populations are simulated, and not real. The homogeneity in neighbourhoods was built in and may be greater than in real life. Still, similarity of nearby households is broadly realistic. We also note that our simulated samples were ideal and ignored logistical difficulties that real surveys experience [16]. These include the fact that population numbers are not exact (so PPS sampling is really PPES), interviewer teams make decisions that may not strictly follow protocols, and people in households may be out when interviewers call or may refuse to participate. Still, we expect such problems would in practice be similar for all or most sampling methods.

Other criteria for comparison

Several criteria apply when comparing sampling procedures [117]. While they were applied to new approaches using Geographic Information System (GIS) technology, most of the criteria are relevant here.

Coverage: The Square method relies on identification of buildings or households from aerial images. There are likely to be errors in such identification, both false positives (features misclassified as buildings) and false negatives (buildings misclassified and thus not included in the possible set of sampled buildings). Other buildings may be missed altogether because of, e.g., tree or cloud cover. Such errors are less likely with on-the-ground survey teams. The EPI methods, though, are likely to omit more isolated homes, unless they are the first household selected, since they will rarely be the next nearest neighbour.

Cost: If the main cost of a survey is travel to the PSU (town or EA), all methods will have similar cost. The Square method may require more travel within towns, which could be quite substantial – and costly - for large towns.

Speed: Several stages of the surveys are common to all methods: obtaining population estimates for the PSUs, conducting interviewing, cleaning and collating data, and conducting the analysis.

For the NewEPI, far more information (data on Enumeration Areas) is needed before the survey can be done. It may be readily available in official files, but access may take time. As well, survey teams must list all households in the PSU, and either select all of them or choose a random sample – this can be done concurrently with the interviews themselves. The NewEPI manual projects a 12-month timetable from conception of a national survey to its completion [6:23], far longer than required in emergencies. The Square method requires obtaining aerial images and identifying the sample from them, a relatively quick process. The older EPI methods require interviewer teams to identify buildings, which is likely to be quicker and it can be combined with the interviews in one trip to a PSU.

Degree of Interviewer Involvement: The OldEPI methods rely on interviewers to identify households in a random direction, randomly select one, and choose the next nearest households. All of these are subject to error. The NewEPI requires survey teams to list all households in an EA. Depending on how clear the boundaries are, error may arise. The Square method depends in part of the precision of the GPS locators and quality of maps, and how readily teams can identify the selected building.

Control over probabilities of selection: The probabilities of selection can be estimated for the GPS methods and the NewEPI, but not for the older EPI methods. The similar size of the EAs leads to similar sampling weights, and smaller standard errors of estimates than methods like the GPS techniques that can have quite different sampling weights depending on the variation in population density.

Technical Skills: Some sources of information ‘require advanced training to use properly’ [15:70]. Since the methods we have described are readily available and easily usable today, this is not a concern. (Some gridded population datasets are an exception.) Moreover, as technologies improve, the accuracy of identifying buildings will also get better.

Conclusions

The simulations show that the Square technique, entailing the use of GPS and satellite images, almost always produces the lowest RMSE when estimating the population prevalence of, e.g., disease or vaccination coverage. The Square method is also generally the best (lowest RMSE) when estimating relative risks. Its overall superiority was unrelated to any of the population characteristics we built into the populations, such as variability in population density across towns. While methods such as NewEPI improve on the original EPI, we recommend the use of the Square methodology, unless there are practical reasons to prefer another technique.

Abbreviations

EA Enumeration Area

EPI Extended Program on Immunization

GPS Global Positioning System

MSE Mean Squared Error

PPES Probability Proportional to Estimated Size

PPS Probability Proportional to Size

PSU Primary Sampling Unit

RMSE Root Mean Squared Error

RR Relative Risk

SRS Simple Random Sampling

WHO World Health Organization

Declarations

Ethics Approval and consent to participate

Since no humans or their data were used, ethical approval was not required.

Consent for publication

Not applicable

Availability of data and materials

Since ours was a simulation study in which we created the populations, we do not have data per se. The software for creating the populations and running the sampling methods was written primarily by one of us (PDE).

- Project home page: <https://github.com/patrickdemond/sampsim/>
- Archived version: The source code can be downloaded at <http://doi.org/10.5281/zenodo.4731010>
- Operating system(s): Platform independent
- Programming language: R, C++
- Other requirements: cmake, and LibArchive, JSON Cpp, and C++11 compliant compiler must be linked for the project to compile
- License: GNU General Public License 3.0. <https://choosealicense.com/licenses/gpl-3.0/>
- Any restrictions to use by non-academics: No

The program output is included in Additional file 4 as an Excel file, together with Additional file 5 on how to interpret the Excel file. The additional files also contain tables not included in the manuscript.

Competing interests

The authors declare that they have no competing interests

Funding

The study was funded by the Canadian Institutes of Health Research, Funding Reference Number: 123432.

Authors' contributions

HSS and RV-A conceived the study, and HSS, RV-A, and BB developed the protocol. PDE programmed and undertook the simulations, with input from BB, HSS, and RV-A. HSS wrote the initial draft of the manuscript. All authors contributed to revisions and read and approved the final manuscript.

Acknowledgements

Not applicable

References

- 1 World Health Organization. Immunization coverage cluster survey – Reference manual. Geneva: World Health Organization; 2005.
- 2 Kolbe AR, Hutson RA, Shannon H, Trzcinski E, Miles B, Levitz N, et al. Mortality, crime and access to basic needs before and after the Haiti earthquake: A random survey of Port-au-Prince households. *Medicine, Conflict and Survival*. 2010;26:281-297.
- 3 Miller AC, Rohloff P, Blake A, Dhaenens E, Shaw L, Tuiz E et al. Feasibility of satellite image and GIS sampling for population representative surveys: a case study from rural Guatemala. *Int J Health Geogr*. 2020;19:56.
- 4 Bennett S, Radalowicz A, Vela V, Tomkins A. A computer simulation of household sampling schemes for health surveys in developing countries. *Int J Epidemiol*. 1994;23:1282-91.
- 5 Brogan D, Flagg EW, Deming M, Waldman R. Increasing the accuracy of the Expanded Programme on Immunisation's cluster survey design. *Ann Epidemiol*. 1994;4:302-11.

- 6 Turner AG, Magnani RJ, Shuaib M. A not quite as quick but much cleaner alternative to the Expanded Programme on Immunisation (EPI) cluster survey design. *Int J Epidemiol*. 1996;25:198-203.
- 7 WHO. Vaccination coverage cluster surveys: Reference manual. Geneva: World Health Organization; 2018.
- 8 Henderson RH, Sundaresan T. Cluster sampling to assess immunization coverage: a review of experience with a simplified sampling method. *Bull World Health Organ*. 1982;60:253-60.
- 9 Lemeshow S, Tserkovnyi AG, Tulloch JL, Dowd JE, Lwanga SK, Keja J. A computer simulation of the EPI survey strategy. *Int J Epidemiol*. 1985;14:473-81.
- 10 Harris DR, Lemeshow S. Evaluation of the EPI survey methodology for estimating Relative Risk. *Wld Hlth Statist Quart* 1991;44:107-114.
- 11 Blower SM, Dowlatabadi H. Sensitivity and uncertainty analysis of complex models of disease transmission: An HIV model, as an example. *International Statistical Review / Revue Internationale de Statistique*. 1994;62:229-243.
- 12 Grais RF, Rose AMC, Guthmann J-P. Don't spin the pen: two alternative methods for second-stage sampling in urban cluster surveys. *Emerg Themes Epidemiol*. 2007;4:8 doi:10.1186/1742-7622-4-8.
- 13 Shannon HS, Hutson R, Kolbe A, Haines T, Stringer B. Choosing a survey sample when data on the population are limited: a method using Global Positioning Systems and satellite photographs. *Emerg Themes Epidemiol*. 2012;9:5.
- 14 Thomson DR, Rhoda, DA, Tatem A, Castro MC. Gridded population survey sampling: a systematic scoping review of the field and strategic research agenda. *Int J Health Geogr*. 2020;19:34.
- 15 Anon. About SHARCNET. <https://www.sharcnet.ca/my/about>. Accessed 14 April 2021.
- 16 Cutts FT, Izurieta HS, Rhoda DA. Measuring coverage in MNCH: design, implementation, and interpretation challenges associated with tracking vaccination coverage using household surveys. *PLoS Med*. 2013;10:e1001404.
- 17 Eckman S, Himelein K, Dever JA. Innovative sample designs using GIS technology. Chapter 4, pp 67-92, in Johnson JP, Pennell B-E, Stoop IAL, Dorer B, editors. *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*. Hoboken, NJ: John Wiley & Sons; 2019.

Figures

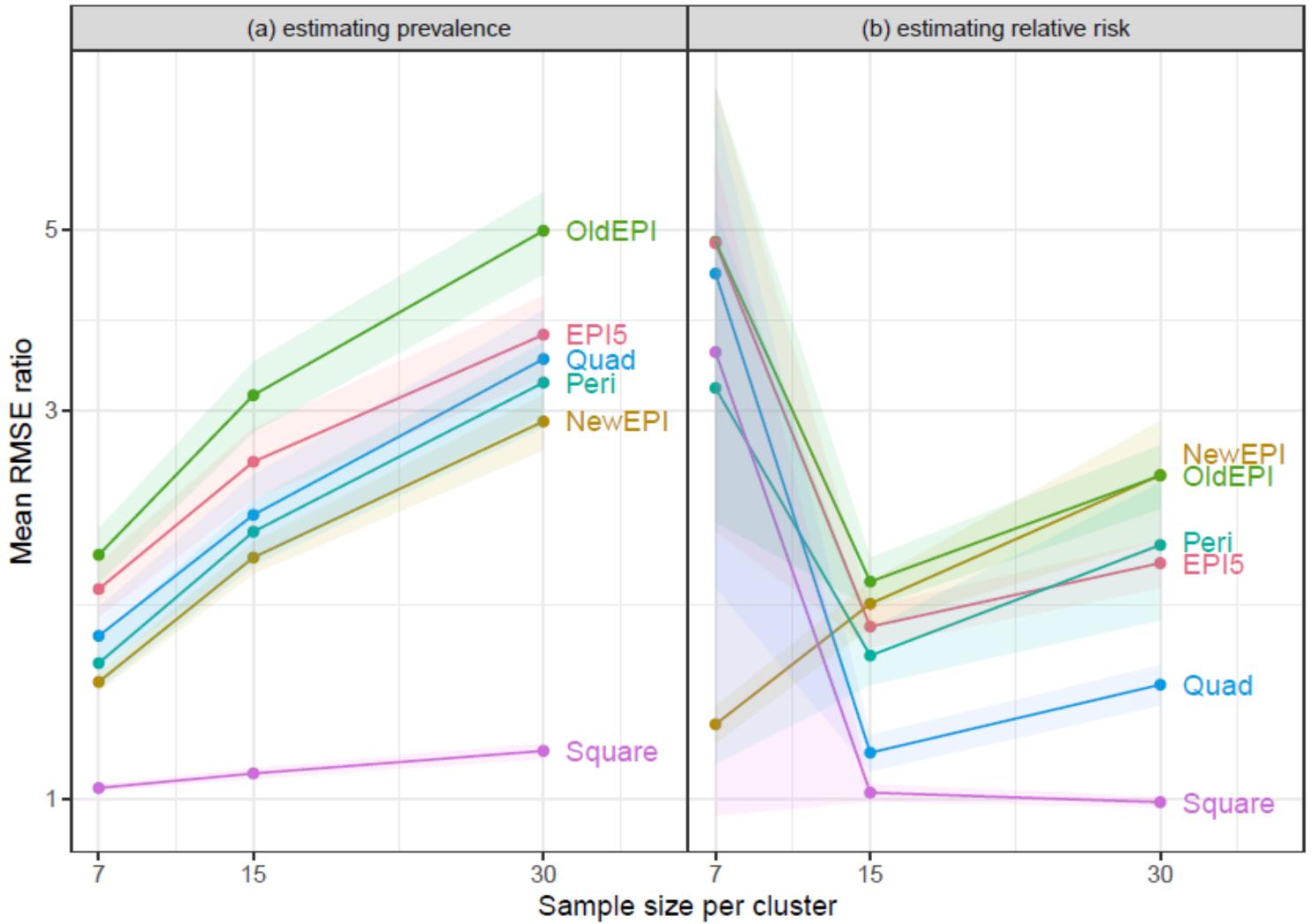


Figure 1

Title: Mean of ratios of RMSE for sampling method to RMSE for simple random sampling. Legend: Figure shows the mean ratios when estimating (a) Prevalence and (b) Relative Risk (RR), using the same sample of towns (clusters for the NewEPI method) for each population, and RR=1.0.

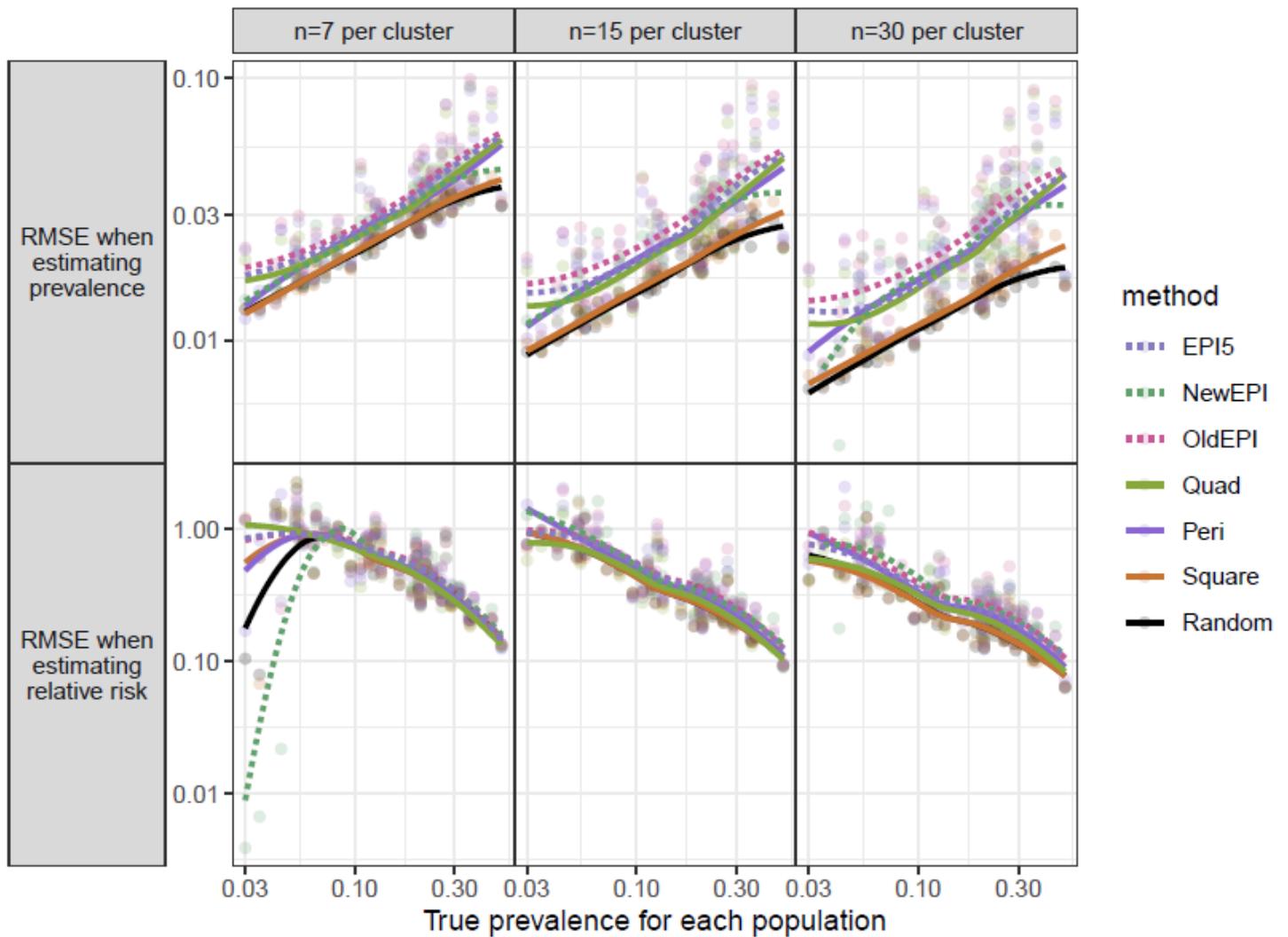


Figure 2

Title: Root Mean Square Error (RMSE) for each population against prevalence Legend: Figure shows RMSE for three sample sizes, when the same towns were sampled for each simulation and relative risk (RR) = 1.0.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1Parameterstable.docx](#)
- [Additionalfile2Diseasedeterminationalgorithm.docx](#)
- [Additionalfile3Fullsetoftables.docx](#)
- [Additionalfile4SimulationOutput.xlsx](#)
- [Additionalfile5HowtointerpretSimulationOutputfile.docx](#)