Population Disease Occurrence Models Using Evolutionary Computation

Jacob Barhak
Jacob Barhak
Austin, TX
jacob.barhak@gmail.com

Aaron Garrett
Wofford College
Spartanburg, SC
garrettal@wofford.edu

Anselm Blumer, Olaf Dammann
Tufts University
Boston, MA
olaf.dammann@tufts.edu
ablumer@cs.tufts.edu

ABSTRACT

The availability of computing power now allows for computation methods that seemed expensive in the past. This enables the exploring of synthetic population characteristics while running simulations at the individual level. Specifically, it is now possible to generate synthetic populations that mimic population statistics of published epidemiological data and to explore hypothetical scenarios. This work shows how Evolutionary Computation (EC) techniques can be used to create a population occurrence model that projects possible treatment effects on population outcomes using characteristics that are intrinsic to the population. We demonstrate how EC is used to extend a previous solution to the population disease occurrence model and generalize it. This exploration reveals the need for epidemiological experts to provide additional information to accompany publication of population statistics to support machine comprehension.

ABOUT THE AUTHORS

Olaf Dammann is Professor of Public Health and Community Medicine at Tufts University School of Medicine. He is interested in the elucidation of perinatal risk scenarios in the context of maternal intrauterine infection and inflammatory responses of the mother, fetus, and newborn. The main outcomes researched are perinatal brain damage and retinopathy of prematurity. He is also interested in theory of risk and causation. Also see: https://medicine.tufts.edu/faculty/olaf-dammann

Anselm Blumer is an Associate Professor Emeritus in Computer Science at Tufts University. His primary research interest is in machine learning, particularly in applications to biology and medicine. He has also done work on data compression and algorithms for indexing and search. Also see: https://engineering.tufts.edu/cs/people/faculty/anselm-blumer

Jacob Barhak specializes in population modeling and specifically in chronic disease modeling with emphasis on using computational technological solutions. Dr. Barhak has diverse international background in engineering and computing science. The Reference Model for disease progression was independently self-developed by Dr. Barhak in 2012. He is the developer of the MIcro Simulation Tool (MIST). See: http://sites.google.com/site/jacobbarhak/

Aaron Garrett is an Assistant Professor in computer science at Wofford College. His interests include evolutionary computation and machine learning. He is the author of INSPYRED, a software library that includes biologically-inspired computation and encompasses a broad range of algorithms including evolutionary computation, swarm intelligence, and neural networks. For additional information please visit http://sites.wofford.edu/garrettal/

Population Disease Occurrence Models Using Evolutionary Computation

Jacob Barhak
Jacob Barhak
Austin, TX
jacob.barhak@gmail.com

Aaron Garrett
Wofford College
Spartanburg, SC
garrettal@wofford.edu

Anselm Blumer, Olaf Dammann
Tufts University
Boston, MA
olaf.dammann@tufts.edu
ablumer@cs.tufts.edu

INTRODUCTION

Moore's law has been driving the exponential growth of computing power for over half a century (Moore, 1965). For the price of one hour of work, it is possible today to buy sufficient computing power to perform computations that were considered unreasonable last century. With this computing power, it is possible to simulate many individual entities within a population or employ time consuming algorithms to better understand the properties of populations.

In turn, the population modeling field has recently advanced from modeling population cohorts to modeling individuals. Numerous examples of such modeling (Population Modeling Working Group, 2015, 2016, 2017) require generation of synthetic populations to model reality. Medical data are typically restricted and therefore less available for researchers, mostly due to privacy issues, yet is widely published as summary data in the literature. Many examples can be found in clinical trials that are now aggregated according to US law in ClinicalTrials.Gov (ClinicalTrials.Gov, Online). The Reference Model (Barhak , 2015) is one example where information from many clinical trials is modeled to fit their published statistics. This is possible due to the MIcro Simulation Tool (MIST) (Barhak, 2013) that can receive population statistics that can contain inclusion and exclusion criteria to handle skewed distributions that are common in clinical trials. For example, MIST can generate an artificial population of individuals that match the mean age and its standard deviation of the population for a trial with a minimum age requirement.

This work extends the scope of population generation to handle more complex constraints that allows extracting indirect information from the synthetic population in order to project the effect of treatment of a population. We name the type of solution Population Disease Occurrence Model (PopDOM).

POPULATION DISEASE OCCURRENCE MODELS OVERVIEW

This model was first suggested by Olaf Dammann who asked for assistance from the population modeling working group. Without losing generality, we will explain the technique using the same problem shown in (Dammann, Chui and Blumer, 2018) where an initial solution appeared.

Known:
In a population of $N=617$ preterm infants, where $P_1=32\%$ are with Sepsis, $P_2=75\%$ get Oxygen, it was observed that $P_3=47\%$ reached the outcome of Retinopathy of Prematurity (ROP) (Chen, 2011). The odds ratios between parameters and outcomes were $O_{12}=Odds(Oxygen,Sepsis) =2.6$, $O_{13} = Odds(ROP,Sepsis) = 2.8$, $O_{23} = Odds(ROP,Oxygen) = 3.6$,

To be solved:
A new treatment is introduced that reduces the occurrence of sepsis from $P_1=32\%$ to a lower value $P_1^* = 16\%$. Assuming that the odds ratios and the oxygen probability represent biological constraints that do not change, what would be the resulting prevalence (percentage) of ROP?

Solution:
To solve this problem, let us first analyze the problem and explain some elements there. Parameters X_i in the problem, are Boolean parameters for each individual, where $i=1..3$. Parameter X_i for individual $k=1..N$ can be either 0 or 1 and we can write it as X_{ik}. The probability of a parameter i is simply defined as $P_i = \#(X_{ik}=1)/N$ where # represents the count operator. The odds ratio of two parameters i,j defined by the following equation:

$$O_{ij}= \#(X_{ik}=1 \ \& \ X_{jk}=1)^* \#(X_{ik}=0 \ \& \ X_{jk}=0) \ / \ (\#(X_{ik}=0 \ \& \ X_{jk}=1)^* \#(X_{ik}=1 \ \& \ X_{jk}=0))$$

Since X_{ik} parameters are Boolean, we can look at the problem from the cohort perspective where group $G_{abc}=\{X_{1k}=a \ \& \ X_{2k}=b \ \& \ X_{3k}=c\}$. So for this 3 parameter problem, there are 8 groups: $G_{000}, G_{001}, G_{010}, G_{011}, G_{100}, G_{101}, G_{110}, G_{111}$. We will denote the size of group G_{abc} as $\#G_{abc}$. We can also write the odds ratios by replacing $\#(X_{1k}=1 \ \& \ X_{2k}=1)$ with $\#(X_{1k}=1 \ \& \ X_{2k}=1 \ \& \ X_{3k}=0 \ or \ X_{1k}=1 \ \& \ X_{2k}=1 \ \& \ X_{3k}=1)$ which is equivalent to $(\#G_{110}+\#G_{111})$. We can do the same for all other elements.

The solution offered in (Dammann, Chui and Blumer, 2018) is to solve the problem by writing equations for the sizes of each group. Let us start by following in the same path. The given known elements account for 7 equations, 3 equations for probabilities, 3 equations for odds ratios, and 1 equation to account for population size.

1. $P_1 = (\#G_{100}+\#G_{101}+\#G_{110}+\#G_{111}) /N$
2. $P_2 = (\#G_{010}+\#G_{011}+\#G_{110}+\#G_{111}) /N$
3. $P_3 = (\#G_{001}+\#G_{101}+\#G_{011}+\#G_{111}) /N$
4. $O_{12} = (\#G_{110}+\#G_{111}) *(\#G_{000}+\#G_{001})/ ((\#G_{010}+\#G_{011})*(\#G_{100}+\#G_{101}))$
5. $O_{13} = (\#G_{101}+\#G_{111}) *(\#G_{000}+\#G_{010})/ ((\#G_{001}+\#G_{011})*(\#G_{100}+\#G_{110}))$
6. $O_{23} = (\#G_{011}+\#G_{111}) *(\#G_{000}+\#G_{100})/ ((\#G_{001}+\#G_{101})*(\#G_{010}+\#G_{110}))$
7. $N = \#G_{000}+\#G_{001}+\#G_{010}+\#G_{011}+\#G_{100}+\#G_{101}+\#G_{110}+\#G_{111}$

Since there are 8 unknown values for $\#G_{abc}$ and 7 equations, it leaves one degree of freedom for a solver to calculate the group counts. In the solution in (Dammann, Chui and Blumer, 2018) another assumption was made to account for the 8th equation to allow a solution for this problem. The added assumption kept constant ratios involving outcome and risk factor group and was added outside the formal definition of the problem. Since the assumption was not based on a given constraint and although may be reasonable, it may not represent reality. In this paper we will avoid this assumption and explore the multiple population distributions possible.

Moreover, the solution given in (Dammann, Chui and Blumer, 2018) has only $D=3$ underlying parameters: Sepsis, Oxygen, ROP and 8 unknowns. The solution space for this problem grows exponentially with the number of parameters – 16 unknowns with 4 parameters and 2^D unknowns for D parameters. With the simple problem given with 3 parameters, it is possible to compute all possible permutations on a modern computer to find all $\#G_{abc}$ combinations that make sense. However, if we have the same problem with more parameters, the number of degrees of freedom will increase and solution may not be practical, especially if some information such an odds ratio is not given. However, a solution that explores distributions may provide better insight on the problem. Therefore, we chose an Evolutionary Computation (EC) solution that can cope with large solution spaces.

POPULATION GENERATION USING EVOLUTIONARY COMPUTATION

The EC solution chosen is a type of Genetic Algorithm. The idea is that the EC solution generates a population of candidate solutions. Each candidate solution is a population of individuals that may solve this problem given. To avoid confusion, note that we generate a population of populations. The quality of the solution is determined by a fitness function that quantifies the error of the solution from the ideal population. In this work the fitness function is of the form:

$$Fitness(s) = W_1\Sigma|P_i - P_i'|+ W_2\Sigma|O_{ij} - O_{ij}'|$$

Where W_1, W_2 are constants and P_i' and O_{ij}' are the probabilities and odds ratios of the candidate solutions. The ideal solution will result in a zero value for the fitness function.

The EC solution walks through these main stages:

1. Generation: A population of random solutions is generated. In this problem, each solution s consists *of* $D \times N$ random numbers $X_{ik} \sim Bernoulli(P_i)$ where $i=1..D$, $k=1..N$ and P_i is the probability given – while odds ratios are initially ignored. These are the initial conditions for the problem.
2. Evaluation: Where $Fitness(s)$ is calculated for each solution s
3. Selection: Where the best solutions are ranked and selected to represent the next generation
4. Variation: Where the selected solutions undergo mutation and crossover operators to create another generation – consisting of a population of solutions. The operators defined in this work are:

   a. Cross-over: from a pair of mother and father solutions $s_1 = \{X_{ik}\}$ and $s_2 = \{Y_{ik}\}$ create two offspring solutions such that $s_3 = \{ X_{ik}$ *if* $R_k=0$, Y_{ik} *if* $R_k=1\}$, $s_4 = \{ Y_{ik}$ *if* $R_k=0$, X_{ik} *if* $R_k=1\}$ where R_k~*Bernoulli(0.5)*
   b. Internal Swap Mutator: swap a single parameter value X_i between two random individuals k and h in the solution so that these two values are swapped: $X_{ik} \Leftrightarrow X_{ih}$. Do this for each individual k if *Uniform(0,1)*< R where R is the mutation rate.
   c. Reroll Mutator: regenerate the individual X_{ik} in solution s, $X_{ik} \sim Bernoulli(P_i)$ where $i=1..D$, $k=1..N$. Do this for each individual k if *Uniform(0,1)*< R where R is the mutation rate.
5. Termination: where a stopping criteria is checked – in this paper the algorithm stops after a certain number of generations is reached. If a stop criteria was not reached, the EC algorithm goes back to step 2
6. Post termination, the most fitting population is considered the answer

Since there is randomness involved in several stages of the EC algorithm, the solution may have random elements. Therefore to get a better understanding, the EC algorithm simulation was repeated several times to show a distribution of results.

However, it is important to point out that the EC algorithm, unlike an analytic solution, will take into account the discrete nature of the problem. In a population of 617 people, there is a finite number of combinations of how groups G_{abc} are formed and each one of those groups will have an integer rather than a real number with a fraction to represent the solution. This type of constraint is hard to solve analytically, yet the EC algorithm incorporates this into its solution and it is likely that the fitness of those solutions will not be the same, so given enough time, it is likely that the discrete nature of the problem will lead to a specific solution – even though many solutions are possible that are very close to the given numbers.

It is also important to note the problem itself is somewhat ill defined since epidemiological studies rarely report exact number with high precision. If high precision numbers would be reported, it would perhaps be possible to find the exact population that matches all the statistics provided since the discrete nature of the problem would point to a specific solution. However, when the numbers given as imports are truncated to a certain precision, it may add some uncertainly if there are several close solutions. However, from the epidemiological perspective, the epidemiological study that created the input numbers has some statistical variation and if repeated, it may not provide the same numbers. Therefore, we are still interested in distribution of close results that the EC solution can provide.

However, so far the focus was on generating a population that matches statistics whereas the problem tries to calculate the effect of treatment. The next section will address this important solution step.

POPULATION DISEASE OCCURRENCE MODELS USING EVOLUTIONARY COMPUTATION

The naive solution for the original problem has two population generation steps:

Step 1: Generate a population that matches the original untreated population statistics using EC.
Step 2: Generate a population with the estimated treatment effect as a constraint while removing the constraint on the outcome.

Table 1. Parameters for the naive solution using two population generation steps.

Solution Step	N	P_1 Sepsis	P_2 Oxygen	P_3 ROP	O_{12} Sep/Oxy	O_{13} Sep/ROP	O_{23} Oxy/ROP
Step 1	617	0.32	0.75	0.47	2.6	2.8	3.6
Step 2	617	0.16	0.75	?	2.6	2.8	3.6

Table 1 shows the parameters to be used in each step of the solution. Note that the question mark in the ROP in step 2 creates some ambiguity since some probability needs to be given to the initial solution generator of the EC. For the purposes for generating the initial population, the step 1 probability is used for ROP. However, there is no constraint on this probability during evolution and it is allowed to drift to get a new value that will match all other constraints.

If we recall the analytic solution with 8 unknowns, we can see that step 1 has one degree of freedom. However, if we try to calculate the population for step 2, there is another degree of freedom added since the ROP outcome is no longer a constraint. Therefore, there are two degrees of freedom. Also note that the naive solution does not have any connections between the steps, in fact step 2 can run before step 1 so step 1 seems to be redundant. Due to the degrees of freedom in solution, it is unclear if the populations generated actually provide a solution. In fact when running simulations this way, some unreasonable results appeared where ROP did not improve as expected due to the two degrees of freedom allowed in step 2.

Therefore, a slight variation was introduced in this solution to pass information between the two steps. To help do that, another new measure was introduced. We call it division ratio. The division ratio between two parameters is defined as R_{ij} = $\#(X_{ik}=1 \ \& \ X_{jk}=1) \ / \ \#(X_{ik}=1 \ \& \ X_{jk}=0)$ which is a subset of the odds ratio calculation. This artificial calculation will help us figure out the missing degrees of freedom that we will transfer between the two steps assuming that those values will remain the same. The population disease occurrence solution will therefore be formulated as follows.

Population Disease Occurrence Model Algorithm

Step 1: Generate a population that matches the original untreated population statistics using EC. Extract invariant properties from the solution.
Step 2: Generate a population with the estimated treatment effect as a constraint while removing the constraint on the outcome and applying the invariant properties as additional constraints.

Table 2. Parameters for the population disease occurrence model

	N	P_1 Sepsis	P_2 Oxygen	P_3 ROP	O_{12} Sep/Oxy	O_{13} Sep/ROP	O_{23} Oxy/ROP	R_{12} Sep/Oxy	R_{13} Sep/ROP	R_{23} Oxy/ROP
Step 1	617	0.32	0.75	0.47	2.6	2.8	3.6	?	?	?
Step 2A	617	0.16	0.75	?	2.6	2.8	3.6	?	?	?
Step 2B	617	0.16	0.75	?	2.6	2.8	3.6	?	?	Step 1
Step 2C	617	0.16	0.75	?	2.6	2.8	3.6	?	Step 1	?
Step 2D	617	0.16	0.75	?	2.6	2.8	3.6	?	Step 1	Step 1
Step 2E	617	0.16	0.75	?	2.6	2.8	3.6	Step 1	?	?
Step 2F	617	0.16	0.75	?	2.6	2.8	3.6	Step 1	?	Step 1
Step 2G	617	0.16	0.75	?	2.6	2.8	3.6	Step 1	Step 1	?
Step 2H	617	0.16	0.75	?	2.6	2.8	3.6	Step 1	Step 1	Step 1

Table 2 shows the parameters to be added as input. The division ratio parameters are not given and need to be extracted from the solution. Therefore step 1 is important since it calculates unseen values that determine how the groups are formed and the division ratios calculated there can be used in step 2. However, the question is which of the 3 division ratios should be used in step 2, in other words, which one actually stays invariant between solutions. Since no information is provided, it was decided to check all possible variations and therefore 8 different step 2 strategies can be executed marked with the letters A-H. In each one of those simulations a different one of those values is transferred from step 1. Since we execute step 1 multiple times, we transfer the mean value calculated from multiple simulations.

Strategy C seems to be the most intuitive one since it ties between the sepsis parameter which was treated to the ROP outcome. However, we also wanted to see what happens when another single division ratio is kept stable and strategies B and E show this. However, note that even if we transfer one invariant from Step 1, there is still a degree of freedom if writing the equations and multiple solutions are theoretically possible. Therefore strategies D,F,G that use extra two division ratios from step 1 are used. We also wanted to see what happens if the problem was over constrained and we transfer all 3 division ratio values from step 1 to step 2 – strategy H handles this situation. Note that strategy A provides a reference of what happens if we do not constrain any of the division ratios.

Since we do not have additional information on the problem, we are exploring different paths to see possible behavior. The division ratios introduced here are artificial constructs. If an epidemiologist could provide another constraint that would be kept between step 1 and step 2, it would help. However, even with these artificial constructs we are able to gain some insight to the problem as the results suggest.

RESULTS

Step 1 of the simulation involves calculating the population as close as possible to the target constraints in Table 1. Table 3 shows the results statistics after repeating the EC calculation 100 times. There was one result that was dominant in 98 simulations marked as Sol 1 and two other results Sol 2 and Sol 3 that appeared once in 100 repetitions that may be considered outliers. The system calculated the average for the odds ratios and passed those to step 2 of the solution.

Table 3. Step 1 Results – numbers are rounded to 3 digits for display

	N	P_1 Sepsis	P_2 Oxy.	P_3 ROP	O_{12} Sep/Oxy	O_{13} Sep/ROP	O_{23} Oxy/ROP	R_{12} Sep/Oxy	R_{13} Sep/ROP	R_{23} Oxy/ROP
Target	617	0.320	0.750	0.470	2.600	2.800	3.600			
Sol 1 (98)	617	0.319	0.750	0.468	2.587	2.798	3.614	6.036	1.775	1.184
Sol 2 (1)	617	0.319	0.749	0.473	2.616	2.804	3.601	6.036	1.814	1.211
Sol 3 (1)	617	0.313	0.752	0.470	2.600	2.793	3.597	6.148	1.797	1.189
Average	617	0.319	0.750	0.468	2.587	2.798	3.614	6.037	1.775	1.184

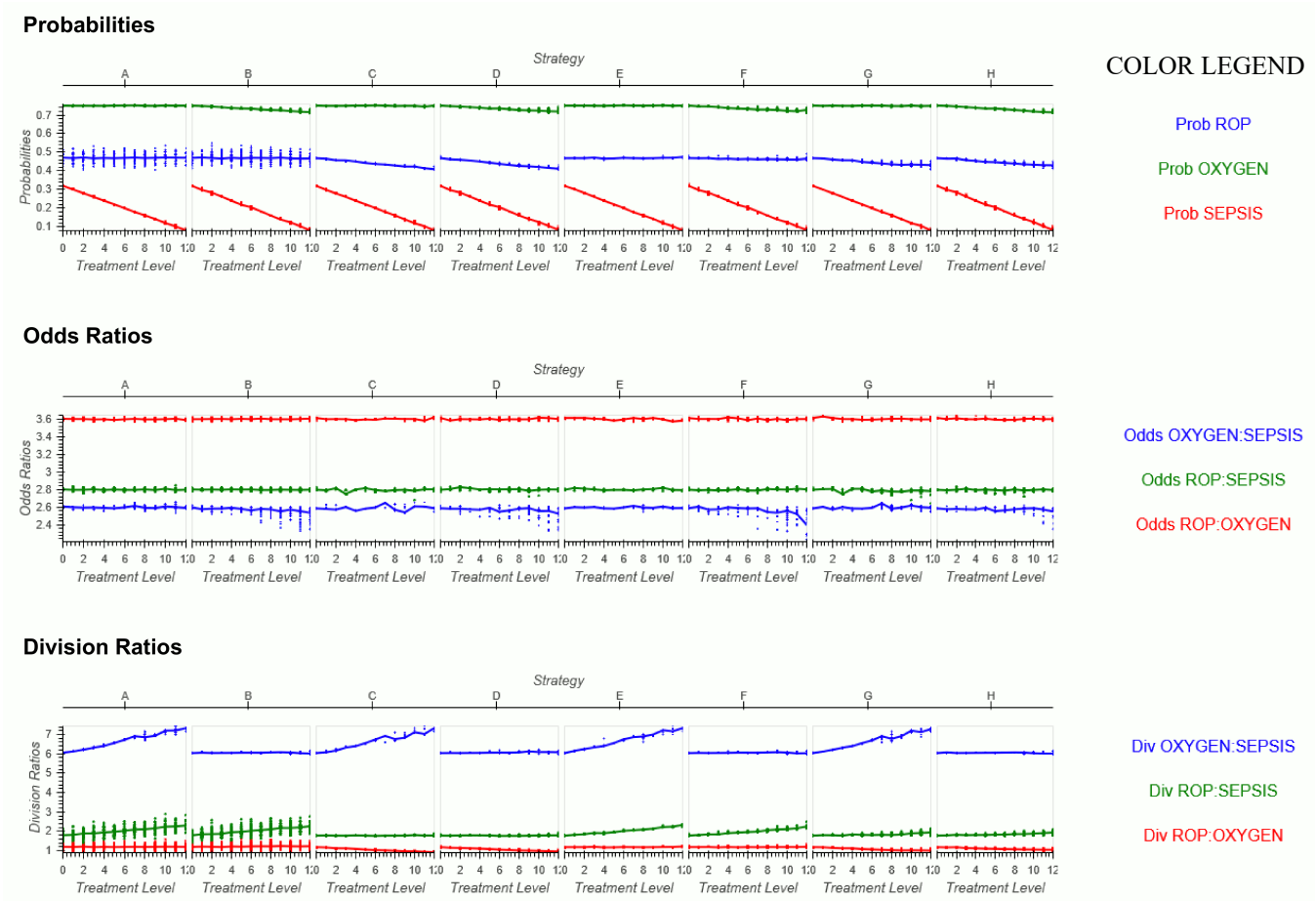

Figure 1. Simulation results for all strategies for different treatment levels. Lines represent average of 100 repetitions while points represent the individual results of the 100 repetitions. The legend on the right represents colors used

The average division ratios highlighted in yellow were passed as parameters to simulation step 2 strategies A-H. The simulation for all strategies A-H was executed for different treatment levels we enumerate from 0 to 12 that reduce sepsis from 0.32 for treatment level 0, to 0.08 in treatment level 12, in steps of 0.02. Figure 1 shows the effect of treatment on output population statistics. The image shows 100 repetitions of simulation for each parameter participating in simulation. The color legend to the right represents the meaning of each series data by color. Solid lines represent average of 100

repetitions while points represent separate repetitions, so variation is visible. The difference between the different categories is clearly visible. Strategies A and B have the most variation of the ROP outcome since the constraints are not associated with the change in sepsis indicated by the red markers that are decreased. Other strategies produce different results. Also note that odds ratios and the oxygen probability were kept more or less steady with a few exceptions in strategies B, D, F, H near the final treatment level. One possible explanation is random behavior, yet those strategies have a constraint on the division ratios of ROP:Oxygen that is part of the odds ratio of the ROP:Oxygen, so the visible variability may indicate the difficulty of the system to comply with all constraints while changing the sepsis – in other words, the more sepsis is lowered, it is harder to find solutions to comply with all constraints and even if there is an optimal solution, the system may need more computational effort to find it so the spread of possible solutions widens. Also note that the more we decrease sepsis, we will eventually reach a limit where it is impossible to find a solution that fulfills all constraints and the system will find the closest solution it can. This is visible looking at strategy H where we would have expected to see parallel lines in all measures except from the oxygen and ROP probabilities – however instead we see a slight drift in most lines. Recall that solution H represent a situation where there are more equations than parameters to compute so the slight variation in all measures can be viewed as the best compromise the system can find to satisfy the over-constrained problem.

To numerically represent results, we extracted the solution for the last treatment level where sepsis was reduced to 0.08 - a quarter of the initial value. Table 4 shows the numeric results, while table 5 shows the deviation of each strategy from the constraints.

Tables 4 and 5 reveal that in most cases, odds ratios changes are minimal while division ratios are more volatile which makes sense due to the formula construct. Each strategy selected seems to have trade off in accuracy of certain constraints while strategy H seems to be spread the inaccuracy between all parameters which makes sense as it is more constrained than other solutions. So unless a medical expert provides some additional criteria for solving the problem in step 2, perhaps the H strategy can be used to give a moderate result.

Table 4. Step 2 Results per strategy – numbers are rounded to 3 digits for display

Strategy	P_1 Sepsis	P_2 Oxygen	P_3 ROP	O_{12} Sep/Oxy	O_{13} Sep/ROP	O_{23} Oxy/ROP	R_{12} Sep/Oxy	R_{13} Sep/ROP	R_{23} Oxy/ROP
Step 1 Ref	0.080	0.750	0.470	2.600	2.800	3.600	6.037	1.775	1.184
A	0.081	0.750	0.470	2.592	2.799	3.590	7.328	2.303	1.198
B	0.081	0.716	0.466	2.537	2.799	3.606	6.015	2.265	1.231
C	0.081	0.750	0.409	2.589	2.803	3.623	7.326	1.782	0.916
D	0.084	0.719	0.412	2.528	2.801	3.601	6.084	1.795	0.962
E	0.081	0.750	0.475	2.590	2.794	3.584	7.331	2.332	1.215
F	0.083	0.728	0.466	2.396	2.803	3.601	6.030	2.248	1.203
G	0.081	0.749	0.431	2.594	2.784	3.597	7.276	1.945	1.010
H	0.084	0.715	0.430	2.553	2.795	3.600	6.010	1.939	1.050

Table 5. Step 2 Deviation from reference per strategy – numbers are rounded to 3 digits for display

Strategy	P_1 Sepsis	P_2 Oxygen	P_3 ROP	O_{12} Sep/Oxy	O_{13} Sep/ROP	O_{23} Oxy/ROP	R_{12} Sep/Oxy	R_{13} Sep/ROP	R_{23} Oxy/ROP
A	0.001	0.000	0.000	-0.008	-0.001	-0.010	1.291	0.527	0.014
B	0.001	-0.034	-0.004	-0.063	-0.001	0.006	-0.022	0.489	0.047
C	0.001	0.000	-0.061	-0.011	0.003	0.023	1.289	0.007	-0.269
D	0.004	-0.031	-0.058	-0.072	0.001	0.001	0.047	0.020	-0.223
E	0.001	0.000	0.005	-0.010	-0.006	-0.016	1.294	0.556	0.030
F	0.003	-0.022	-0.004	-0.204	0.003	0.001	-0.007	0.473	0.019
G	0.001	-0.001	-0.039	-0.006	-0.016	-0.003	1.240	0.170	-0.174
H	0.004	-0.035	-0.040	-0.047	-0.005	0.000	-0.027	0.164	-0.135

A medical expert may immediately look at the ROP column, marked in yellow, to see how much an outcome is improved when a modeled treatment is assumed to drop sepsis to a quarter of the original probability. However, the variability displayed between the different strategies shows a large variation. Note that solution E actually shows increase in ROP which may be counter intuitive, yet looking at the constraints of strategy E we see that the division ratio connecting sepsis to ROP is not constrained – in fact in all strategies that this constraint is not used, the change is ROP is very small. So the medical expert will have to make a judgment call if the connection between sepsis on ROP is constrained by both odds ratios and division ratios. In absence of knowledge the answer of the system may be written as the following sentence:

When modeling the effect of a hypothetical treatment that drops sepsis from 32% to 8% of the population while keeping odds ratio constraints, different models show a change in ROP from 47% to the range of (40.9% - 47.5%) where the most informed model reached 43%.

This solution should be compared to the solution in the first table in (Dammann, Chui and Blumer, 2018), where for Sepsis = 5-10% results in ROP of 40-41%. Our EC solution of 43% seems conservative, although strategy C seems to match in both solutions. Our EC solution, however, shows a variety of solutions that the epidemiologist should consider.

DISCUSSION

The solution provided in this paper is by no way optimal and can be improved in many ways. While still using EC algorithms it is possible to rewrite the solution definition so that instead of generating individuals, the solution will generate Group sizes *#Gabc* and by this speed up the solution considerably. This solution will still have the advantages of the EC solution, yet it will not have the advantage of being general enough to allow more complicated population generation. For example if another parameter would be added to the problem that is not Boolean, e.g. duration of oxygen that would be represented by a real number, it would not be possible to represent it using number of individuals per group. However, with the current solution that generates a population, it would be easy to randomly generate it and add constraints. So the solution provided, although not optimal, opens future opportunities and can be merged with other population techniques (Barhak, 2015).

Although other solutions still have value and can be merged with the techniques presented here, it is important to note that this type of population modeling problem does not require a quick solution. Instead, it requires exploration of the possible solutions to better understand the observed situation. Specifically, in this work it was shown that the problem is not fully defined and epidemiological experts should provide additional information when published to properly represent the population studies. The following is recommended:

1. Report statistics with higher precision. This practice is considered unpopular within the biomedical community where the statistical error is many times larger than the precision of one digit. However, if computers are asked to reconstruct population characteristic of a population, those numbers can prove useful as those provide ways to check solutions. An appendix to a paper publishing results with the full precision numbers would also help in reproducibility and verification modeling issues. To support this claim, see the calculations in the appendix in (Hayes et. al. 2013) to show how verification can fail when insufficient digits of precision are given.
2. Provide additional measurements of a population. Even clinical trial reports that attempt to provide information on the structure of a population tend to limit information. Even where reporting statistics to ClinicalTrials.Gov (ClinicalTrials.gov, Online), which is now required by law (110th Congress, 2007), studies under report information and many times stick to the very basic Age and Gender categories and do not match data provided in publications. If clinicians want to have their study results reproduced virtually, they have to provide sufficient information for the data to be machine comprehensible.
3. Alongside known facts, clinicians should provide possible explanations even if measurements are not reported. Algorithms such as EC algorithms can get additional possible explanations and see how those fit the reality. Specifically in this problem, if the clinician would define additional possible biological invariants it would be possible to verify those and provide a more conclusive answer. For example, if one of the added constraints to the second step would be defined as invariant, it would be possible to verify if indeed it is invariant.

ACKNOWLEDGEMENTS

The authors acknowledge the help of many open source developers that created free tools to enable this work, and more specifically, to developers of Anaconda, bokeh, dask, and holoviews. Special thanks to Matthew Rocklin who provided insight into the dask library and James Bednar who provided a friendly introduction to holoviews.

REPRODUCIBILITY INFORMATION

The results for this paper were calculated on a 4 core laptop computer with Windows 10 deployed by Anaconda (64-bit) with python 2.7.14, dask 0.17.2, bokeh 0.13.0, inspyred 1.0, numpy 1.14.2 , holoviews 1.10.7 and on a 64 core server with Linux 18.04 with Anaconda (64-bit) python 2.7.15, dask 0.19.1, bokeh 0.13.0, inspyred 1.0, numpy 1.15.3, holoviews 1.10.7. The code is stored in the GitHub repository: https://github.com/Jacob-Barhak/PopDOM

The numbers used in this paper are taken from (Dammann, Chui and Blumer, 2018). Those numbers are close to the numbers in (Chen, 2011), yet are not an exact match, so the analysis in this paper should not be considered for epidemiological use without further exploration into the differences.

REFERENCES

Moore. G. E., (1965). Cramming more components onto integrated circuits. Electronics. https://drive.google.com/file/d/0By83v5TWkGjvQkpBcXJKT1I1TTA/view

Olaf Dammann, Kenneth Chui, Anselm Blumer, (2018) A Causally Naïve and Rigid Population Model of Disease Occurrence Given Two Non-Independent Risk Factors, Online Journal of Public Health Informatics, https://doi.org/10.5210/ojphi.v10i2.9357

Chen M, Çitil A, McCabe F, Leicht, Fiascone, Dammann C.E.L., Dammann O., (2011). Infection, oxygen, and immaturity: interacting risk factors for retinopathy of prematurity. Neonatology. 99, 125-32. PubMed https://doi.org/10.1159/000312821

Population Modeling Working Group, Population Modeling by Examples (WIP) (2015) - SpringSim 2015 , April 12 - 15, Alexandria, VA, USA . http://dl.acm.org/citation.cfm?id=2887741

Population Modeling Working Group, Population Modeling by Examples II (2016) - SummerSim 2016 , July 24 - 27, Montreal, CA. https://doi.org/10.22360/SummerSim.2016.SCSC.060

Population Modeling Working Group, Population Modeling by Examples III (2017) - SummerSim 2017 , July 9 - 12, Bellevue, WA, USA. https://doi.org/10.22360/SummerSim.2017.SCSC.013

ClinicalTrials.gov (Online), ClinicalTrials.gov is a database of privately and publicly funded clinical studies conducted around the world. Online: https://clinicaltrials.gov/ accessed 11/16/2019

Barhak J. (2015). The Reference Model uses Object Oriented Population Generation. SummerSim 2015. Chicago IL, USA. Paper retrieved from: http://dl.acm.org/citation.cfm?id=2874946 Presentation retrieved from: http://sites.google.com/site/jacobbarhak/home/SummerSim2015_Upload_2015_07_26.pptx

Barhak J. (2013), MIST: Micro-Simulation Tool to Support Disease Modeling. SciPy, 2013, Bioinformatics track, https://github.com/scipy/scipy2013_talks/tree/master/talks/jacob_barhak Video retrieved from: http://www.youtube.com/watch?v=AD896WakR94

Hayes A.J., Leal J., Gray A.M., Holman R.R., & Clarke P.M. 2013. "UKPDS outcomes model 2: a new version of a model to simulate lifetime health outcomes of patients with type 2 diabetes mellitus using data from the 30 year United Kingdom Prospective Diabetes Study: UKPDS 82". *Diabetologia*, Vol. 56(9), pp. 1925-33. http://dx.doi.org/10.1007/s00125-013-2940-y

110th Congress, PUBLIC LAW 110–85—SEPT. 27, 2007 (2007) - TITLE VIII—CLINICAL TRIAL DATABASES . Section 801 of the Food and Drug Administration Amendments Act of 2007. Online: https://www.gpo.gov/fdsys/pkg/PLAW-110publ85/pdf/PLAW-110publ85.pdf#page=82