A Causally Naïve and Rigid Population Model of Disease Occurrence Given Two Non-Independent Risk Factors

Olaf Dammann,[1,2] Kenneth Chui, [1] and Anselm Blumer, [2]

1. Department of Public Health and Community Medicine, Tufts University School of Medicine, Boston, MA

2. Department of Gynecology and Obstetrics, Hannover Medical School, 30623 Hannover, Germany

3. Department of Computer Science, Tufts University School of Engineering, Tufts University, Medford, MA

ABSTRACT

We describe a computational population model with two risk factors and one outcome variable in which the prevalence (%) of all three variables, the association between each risk factor and the disease, as well as the association between the two risk factors is the input. We briefly describe three examples: retinopathy of prematurity, diabetes in Panama, and smoking and obesity as risk factors for diabetes. We describe and discuss the simulation results in these three scenarios including how the published information is used as input and how changes in risk factor prevalence changes outcome prevalence.

DOI: 10.5210/ojphi.v10i2.9357

1. Introduction

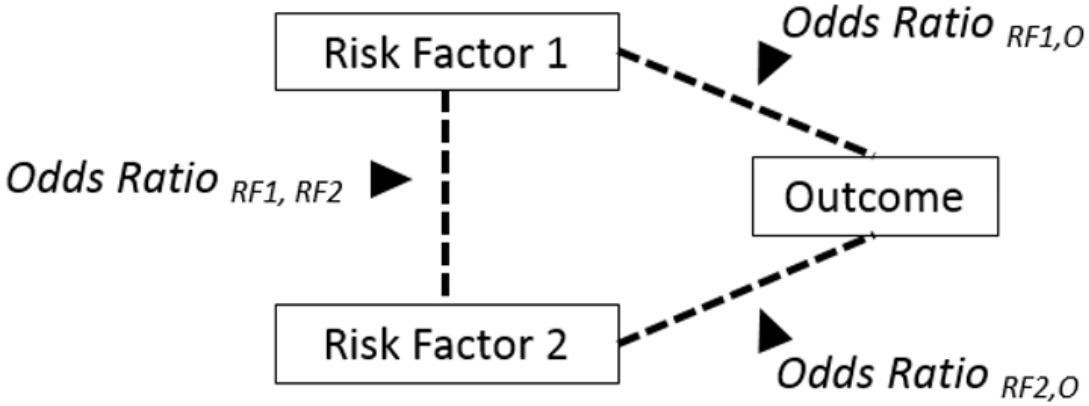
In epidemiology, the concept of multi-causality holds that the occurrence of any disease depends on a set of risk factors, not just one. The generation of virtual databases that reflect the properties of populations is called *micro-simulation* [1]. In their simplest form, such models require as input two risk factors and their association with one outcome variable.

One example is SYNTHEA, a virtual population of individuals and their electronic health records (EHRs) [2]. The algorithm could simulate individuals with, say, three characteristics: a binary disease outcome (coded as yes/no) and two binary risk factors (yes/no). The algorithm uses as input parameters the population prevalence of the two risk factors and the outcome variable; the allocation of “yes” or “no” for each variable is done by applying a Monte-Carlo simulation that uses random numbers and the population prevalence as a threshold. This ensures that, for instance,

on average 37% of the virtual population will have a certain disease if the real population prevalence of that disease is 37% and the threshold for "disease = yes" is set at 0.37.

These microsimulations have one particular disadvantage: if the presence or absence of each variable in the final database is based on separate yes/no attribution processes, the variables will be independent. This, of course, is highly unlikely in reality, because the very definition of a risk factor is that it is associated with the disease under investigation. Moreover, the two risk factors will be independent of each other, which is also rarely the case in real life situations. This way of performing microsimulations will lead to populations that look like their real-life counterparts only with regard to the population average of risk factors and outcome. However, these datasets cannot be utilized to simulate population-wide changes in risk factors with the goal to study population-wide changes in the outcome (disease). Therefore, we wanted to design a model that requires as input the population prevalence of the outcome of interest and of two risk factors, as well their three associations (Figure 1).

Figure 1. The associations among two non-independent risk factors and one outcome are quantified by three odds ratios.



In what follows, we describe a population model with two risk factors and one outcome variable in which the prevalence (%) of all three variables, the association between each risk factor and the disease, as well as the association between the two risk factors is the input. We briefly describe three examples: (#1) retinopathy of prematurity; (#2) diabetes in Panama, and (#3) smoking and obesity as risk factors for diabetes. Next, we describe the simulation results in these three scenarios including how the published information is used as input (Step 1) and how changes in risk factor prevalence changes outcome prevalence (Step 2).

2. METHODS

2.1 The Model

Suppose we have a standard 2 x 2 table for an outcome against a risk factor (Figure 2). Label the cells A, B, C, D where A is the percent of the population for which both the risk factor and the outcome are positive, B is the percent where the risk factor is positive but the outcome is negative, C is the percent where the risk factor is negative but the outcome is positive, and D is the percent where both are negative. Then if RF is the percent of the population with positive risk factor and OUT is the percent of the population with positive outcome, we have

(1) B = RF - A
(2) C = OUT - A
(3) D = 100 - A - B - C

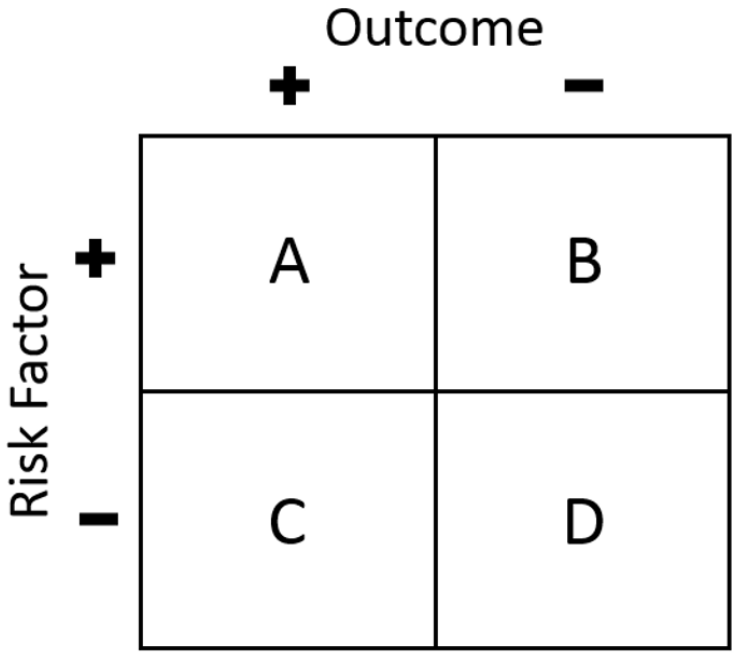
The equation for the odds ratio is based on the quantities depicted in Figure 2:

(4) OR = AD/BC.

We can substitute for B, C, and D using the first three equations, giving a quadratic equation for A with coefficients in terms of RF, OUT, and OR:

$$(5)\ (OR-1)A^2 + (100+(OR-1)(RF+OUT))A + OR \bullet RF \bullet OUT = 0$$

Figure 2. Fourfold table depicting the four entities defined by the presence (+) or absence (-) of a binary risk factor and an outcome.



Solving this will give a 2 x 2 table that matches the given population values for RF and OUT and has the desired odds ratio. This much is calculated in "Step 1" in the JavaScript implementation of the model (available at http://www.cs.tufts.edu/~ablumer/PopStat.html).

We can also use this equation to model the effect of keeping the odds ratio fixed and changing the percentage of the population that has the risk factor.

This can be done by replacing A and RF in the above equation with r*A and r*RF and solving for the value of OUT that keeps the odds ratio constant. This assumes that relative percentages of the population with positive and negative outcome within positive risk factor (A relative to B) stay the same when the positive risk factor population is changed. Since we have two risk factors, we can do identical calculations relating risk factor 1 to the outcome and relating risk factor 2 to the outcome. Similarly, we can find the entries for the 2 x 2 table relating risk factor 1 to risk factor 2.

2.2 Examples

2.2.1 Example #1: Retinopathy of prematurity

We previously analyzed a data set of 617 very preterm newborns [3]. In that project, we found that 47% of all babies developed retinopathy of prematurity (ROP), a serious eye disorder among extremely preterm infants [4]. Systemic inflammation [5] and oxygen exposure data [6] are competing pathogenetic mechanisms that interfere with normal vasculogenesis [7]. The capability to simulate interventions on one or both of these pathomechanisms in order to study changes in ROP occurrence would be a groundbreaking step towards the prevention.

In our data analysis, we also found that 32% of the infants had sepsis and 75% had been exposed to high levels of oxygen. The association between sepsis and oxygen on the one hand and ROP on the other (measured as an odds ratio, OR) were 2.8 and 3.6, respectively. The OR for the association between sepsis and oxygen was 2.6. In Figure 3 we clarify how these data were then entered into the model.

2.2.2 Example #2: Diabetes in Panama

A second example is a study on diabetes in Panama (5.4%) [8] with female sex (RF1: 60%) and age 50+ years (RF2: 31%) as risk factor exemplars. Female sex was associated with diabetes with an OR=1.4, age 50+ had an OR=5.1. The OR for the association between female sex and age 50+ was 0.85 (see Figure 4).

Obviously, in this case, the risk factors are not to be modified to simulate a population intervention as in the previous example. Instead, we are interested in the effect on diabetes prevalence due to the discrepancy between the observed age distribution described in [8] (50+ years = 31%) compared to national data published by the United Nations (20%) [9].

Figure 3. Simulation results of Step 1 in example #1, retinopathy of prematurity.

Population statistics with two risk factors

Enter percentages and odds ratios in the fields below:

Percentages		Odds ratios	
Risk factor 1	32	RF1/OUT	2.8
Risk factor 2	75	RF2/OUT	3.6
Outcome	47	RF1/RF2	2.6

Now click the "Step 1" button:

Step 1 | Reset to zeroes | Reset to test values

Population percentages:

	OUT +	OUT -
RF1 +	20.51514496	11.48485503
RF1 -	26.48485503	41.51514496

	OUT +	OUT -
RF2 +	40.78197836	34.21802163
RF2 -	6.218021632	18.78197836

	RF2 +	RF2 -
RF1 +	27.45634851	4.543651480
RF1 -	47.54365148	20.45634851

The buttons below will now allow you to increment and decrement RF1 and RF2 by approximately 1/100 for each click, or set either to a new value. The odds ratios stay fixed and other values are adjusted so that the odds ratios are still accurate. Since an odds ratio is given between RF1 and RF2, changing one will change the other (as well as OUT).

Increment RF1 | Decrement RF1 | Increment RF2 | Decrement RF2 | Set RF1 to --> 0 <-- Set RF2 to

2.2.3 Example #3: Smoking, BMI, and Diabetes

A randomized controlled trial (RCT) of estrogen plus progestin (EP) versus placebo was conducted in the 1990s to explore the effect of EP on subsequent development of coronary heart disease (CHD) in postmenopausal women [10]. We wanted to use the publicly available data from this RCT to explore the influence of smoking and body mass on diabetes, and use these data as input for a simulation of the effect of two interventions, smoking cessation weight reduction, on diabetes occurrence.

3. Results

3.1 Example #1

In Step 1, we entered the population percentages for both risk factors and the outcome, as well as the three associations among them. The estimated four-fold tables provided by the model are depicted in Figure 3.

In Step 2, we proceeded to the simulation of risk factor modification.

First, we reduced RF1 incrementally down from 32% to 0% (Table 1). This resulted in a drop of RF2 from 75% down to 70% and a reduction in outcome occurrence from 47% down to 39%.

Table 1. *Example #1.* Risk factor (RF)2 and outcome (OUT) changes when RF1 declines (%).

RF1 (Sepsis)	RF2 (Oxygen)	Outcome (Retinopathy of Prematurity)
32	75	47
30	75	46
25	74	45
20	73	44
15	72	43
10	72	41
5	71	40
0	70	39

Second, we reduced RF2 incrementally down from 75% to 0%. This resulted in a drop of RF1 from 32% down to 18% and a reduction in outcome occurrence from 47% down to 25%.

Third, we calculated that even if both RF were reduced to 0, we are still left with a 21% outcome rate, which is probably attributable to other risk factors.

It is also possible that the odds ratios change as the population statistics approach the extremes.

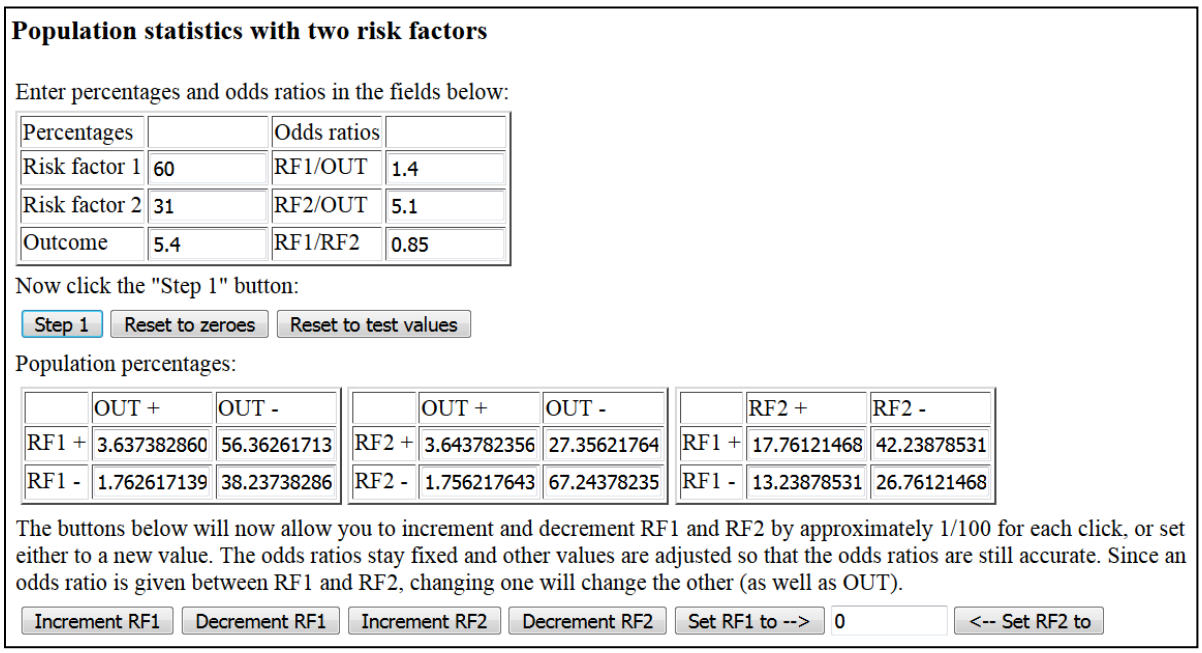
3.2 Example #2

The estimated four-fold tables provided by the model after Step 1 are depicted in Figure 4.

3.3 Example #3

In the publicly available HERS dataset (http://www.biostat.ucsf.edu/vgsm/data.html), we looked at diabetes (on oral medication or insulin) as the outcome, and at smoking and overweight/obesity as risk factors (Table 3). In an exploratory data analysis we found that in this cohort of postmenopausal women with an average age of 67 years, 26% had diabetes, 13% were smokers, and 34% were obese (defined as a BMI ≥30). Smoking was associated with a reduced risk for diabetes (OR 0.5, 95%CI 0.4, 0.7), obesity with a strong risk increase (3.3; 2.7, 3.9), and smoking had an inverse association with obesity (0.6; 0.4, 0.7)(Table 3).

Figure 4. Simulation results of Step 1 in example #2, diabetes in Panama.



Population statistics with two risk factors

Enter percentages and odds ratios in the fields below:

Percentages		Odds ratios	
Risk factor 1	60	RF1/OUT	1.4
Risk factor 2	31	RF2/OUT	5.1
Outcome	5.4	RF1/RF2	0.85

Now click the "Step 1" button:

Step 1 | Reset to zeroes | Reset to test values

Population percentages:

	OUT +	OUT -
RF1 +	3.637382860	56.36261713
RF1 -	1.762617139	38.23738286

	OUT +	OUT -
RF2 +	3.643782356	27.35621764
RF2 -	1.756217643	67.24378235

	RF2 +	RF2 -
RF1 +	17.76121468	42.23878531
RF1 -	13.23878531	26.76121468

The buttons below will now allow you to increment and decrement RF1 and RF2 by approximately 1/100 for each click, or set either to a new value. The odds ratios stay fixed and other values are adjusted so that the odds ratios are still accurate. Since an odds ratio is given between RF1 and RF2, changing one will change the other (as well as OUT).

Increment RF1 | Decrement RF1 | Increment RF2 | Decrement RF2 | Set RF1 to --> 0 <-- Set RF2 to

In Step 2, risk factor modification simulation for Age 50+ from the observed 31% down to the 20% estimated by the UN in a population prevalence decrease for diabetes from 5.4% to 4.4% (data not shown).

Table 2. *Example #1.* Risk factor (RF)1 and outcome (OUT) changes when RF2 declines (%).

RF1 (Sepsis)	**RF2 (Oxygen)**	**Outcome (Retinopathy of Prematurity)**
32	75	47
31	70	46
29	60	43
27	50	40
26	40	37
24	30	34
22	20	31
20	10	28
18	0	25

Table 3. Diabetes among 2758 postmenopausal women, the association between risk factors (smoking and overweight/obese) and diabetes, and the association between risk factors. These data served as input for example #3.

	Diabetes		
	YES	**NO**	**OR (95%C.I.)**
N (row %)	728 (26)	2030 (74)	
Smoking, N (col %)	60 (8)	299 (15)	0.5 (0.4, 0.7)
Obese, N (col %)	397 (55)	545 (27)	3.3 (2.7, 3.9)
Association RF1/RF2	**Smoking**		
N (row %)	**YES**	**NO**	0.6 (0.4, 0.7)
Obese (BMI ≥30), N (col %)	359	2399	
	85 (24)	857 (36)	

We then simulated two interventions, smoking cessation and weight reduction. We have to keep in mind that while obesity is associated with a risk *increase*, smoking is associated with a *decreased* risk for diabetes. The fact that the two risk factors are negatively associated (less obesity among smokers) might explain this "protective effect of smoking".

Reducing smoking to zero in this population led to a minuscule increase of diabetes occurrence from 18 to 19%, which we confirmed in a stratified analysis excluding smokers (Table 4). Among non-smokers, diabetes prevalence was 19.2%.

Reducing obesity was associated with a prominent risk reduction for diabetes, from 18% down to 10%. At the same time, smoking increased from 13 to 17% (Table 5).

Table 4. *Example #3.* Risk factor (RF)2 and outcome (OUT) changes when RF1 declines (%), simulating smoking cessation intervention.

RF1 (Smoking)	RF2 (Overweight/Obesity)	Outcome (Diabetes)
13	56	18
10	57	18
8	57	18
6	57	19
4	58	19
2	58	19
0	58	19

Table 5. *Example #3.* Risk factor (RF)1 and outcome (OUT) changes when RF2 declines (%), simulating weight reduction intervention.

RF1 (Smoking)	RF2 (Overweight/Obesity)	Outcome (Diabetes)
13	56	18
13	50	17
14	40	16
15	30	14
16	20	13
17	10	11
17	0	10

5. DISCUSSION

5.1 Advantages

Our model has three prominent advantages. First, it is novel. To our knowledge, no other population model exists that appreciates the association between risk factors. Second, the model is relatively simple. With only one outcome and two risk factors, the complexity of inputs is limited to their population prevalence and associations between each other. We are currently developing a tool is that includes a third risk factor and that can be used for microsimulations, i.e., it outputs a data file of a virtual population, which can be used in further simulations. Third, the model is freely available online for the community to use and explore.

5.2 Drawbacks

The model is currently limited to two-level exposures and outcomes. It is also limited to only two risk factors. We are currently developing a similar model for three predictors and their inter-relations.

Perhaps the most prominent limitation of the model is that it is causally naïve and rigid. Much of the complex methodology toolbox of modern epidemiology is geared towards the identification of causal risk factors [11]. Our model is not helpful in this regard. The association between risk factors and outcomes is modeled as odds ratios, which are simple measures of strength of association without implying causality or causal direction. The model is also rigid in that the input is reduced to population prevalence and association measures (odds ratios). Within the constraints of these values, the output is not probabilistic, but determined. However, the model can be run multiple times with different values for odds ratios as input that come from within the range of odds ratios defined by the observed confidence interval.

5.3. Conclusion

In this paper, we present a simple model of disease occurrence in populations. Based on the prevalence of a disease and of two risk factors, and of their association with the disease and between each other, the model calculates fourfold tables for these associations (Step 1). Thereafter, the population prevalence of either risk factor can be modified to simulate population risk factor increases or decreases, and changes in disease occurrence can be observed (Step 2). We will now develop this model further to include three risk factors and microsimulation capabilities. In the meantime, we hope it will be helpful to others and would appreciate feedback, preferably in the form of constructive criticism.

Acknowledgements

The following colleagues have contributed to the development of earlier versions of this model: Benjamin Hescott, Inbar Fried, Sadchla Mathieu, Ryan Durgham, Yaa Konama Pokuaa, and Eva Chege. We acknowledge internal support from the TUFTS-Collaborates! Initiative 2014 and Tufts University School of Medicine Chairs' Initiative for Strategic Research Collaborations 2016

References

1. Rutter CM, Zaslavsky AM, Feuer EJ. 2011. Dynamic microsimulation models for health outcomes: a review. *Med Decis Making*. 31(1), 10-18. PubMed https://doi.org/10.1177/0272989X10369005

2. Walonoski J, et al. 2017. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *J Am Med Inform Assoc*. PubMed

3. Chen ML, et al. 2011. Infection, oxygen, and immaturity: interacting risk factors for retinopathy of prematurity. *Neonatology*. 99, 125-32. PubMed https://doi.org/10.1159/000312821

4. Hellstrom A, Smith LE, Dammann O. 2013. Retinopathy of prematurity. *Lancet*. 382(9902), 1445-57. PubMed https://doi.org/10.1016/S0140-6736(13)60178-6

5. Holm M, et al. 2017. Systemic Inflammation-Associated Proteins and Retinopathy of Prematurity in Infants Born Before the 28th Week of Gestation. *Invest Ophthalmol Vis Sci*. 58, 6419-28. PubMed https://doi.org/10.1167/iovs.17-21931

6. Hauspurg AK, et al. 2011. Blood gases and retinopathy of prematurity: the ELGAN Study. *Neonatology*. 99(2), 104-11. PubMed https://doi.org/10.1159/000308454

7. Rivera JC, et al. 2017. Retinopathy of prematurity: inflammation, choroidal degeneration, and novel promising therapeutic strategies. *J Neuroinflammation*. 14(1), 165. PubMed https://doi.org/10.1186/s12974-017-0943-1

8. Mc Donald Posso AJ, et al. 2015. Diabetes in Panama: Epidemiology, Risk Factors, and Clinical Management. *Ann Glob Health*. 81(6), 754-64. PubMed https://doi.org/10.1016/j.aogh.2015.12.014

9. Nations U. *World Population Prospects: The 2017 Revision, DVD Edition*, P.D. Department of Economic and Social Affairs, Editor. 2017.

10. Hulley S, et al. 1998. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Heart and Estrogen/progestin Replacement Study (HERS) Research Group. *JAMA*. 280(7), 605-13. PubMed https://doi.org/10.1001/jama.280.7.605

11. Glass TA, et al. 2013. Causal inference in public health. *Annu Rev Public Health*. 34, 61-75. PubMed https://doi.org/10.1146/annurev-publhealth-031811-124606