Using the copula method to accurately predict the impact of body mass index and alcohol consumption on liver disease

Daniel Ohrenstein[1], Tim Coker[1] Joshua Card Gowers[1], Laura Webber[1], Hannah Graff[1], Juan Vesga[1], Maria Buti[2], Helena Cortez-Pinto[3], Peter Jepsen[4], Jeffrey V. Lazarus[5], Francesco Negro[6], Pierre Nahon[7], Nick Sheron[8], Marietav Simonova[9], Shira Zelber[10], Lise Retat[1]

HEALTH LUMEN

ILC MR 2021

Introduction

Non-communicable diseases (NCDs) such as diabetes, chronic liver disease (CLD) and cancer are the leading cause of morbidity and mortality globally. Increasing rates of NCDs put economic strain on health systems and wider society. Accurately projecting the incidence of NCDs requires the joint prevalence distributions of several relevant risk factors. For example, body mass index (BMI) and air pollution are risk factors for type 2 diabetes; alcohol consumption and BMI are risk factors for CLD; smoking and air pollution are risk factors for lung cancer. However, in most cases, relative risks for diseases have been calculated for individual risk factors only, with joint relative risks for multiple risk factors available only in a small number of cases.

Study Aim : To introduce an innovative and simple implementation approximation called the copula method (often deployed in the financial domain) for estimating joint risks.

Methods

Joint risk factor prevalence information is most accurately estimated from longitudinal studies. Since these studies are rare, we appeal to approximate methods in order to estimate joint risk factor prevalence from one-dimensional risk factor prevalence. It is frequently the case that one-dimensional risk factor prevalence (e.g. proportions of the population in various alcohol consumption groups (irrespective of obesity)) are known. The copula method uses correlation information and the one-dimensional risk factor prevalence to generate an estimate of the joint risk factor prevalence **(Figure 1** and **Table 1).** A general mathematical derivation of the copula method can be found in the literature [1]. We applied this method in the public health domain.

Figure 1. Structural outline for using the copula method to estimate joint relative risks for two risk factors. There are four input prevalence required and two correlations. The copula method is used twice to combine these into two separate joint prevalence for the disease group and the general population. The final step is to take the ratio of these joint prevalence giving the joint risk ratios.

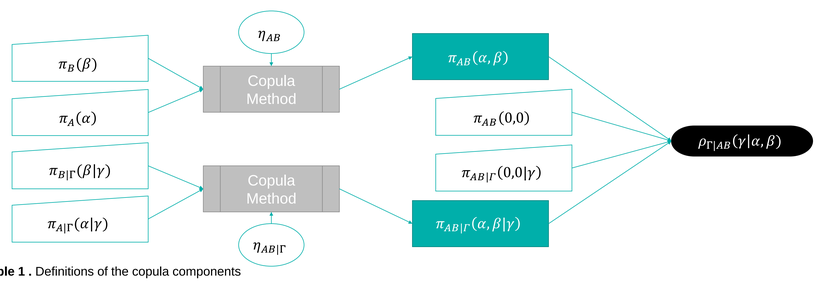


Table 1 . Definitions of the copula components

SYMBOL	DESCRIPTION
$\boldsymbol{\pi}$	Denotes a probability distribution
$\boldsymbol{\pi_A(\alpha)}$	Probability that risk factor A is in state α
$\boldsymbol{\pi_{AB}(\alpha,\beta)}$	Probability that risk factors A and B are in the states α and β respectively
$\boldsymbol{\pi_{AB\|\Gamma}(\alpha,\beta\|\gamma)}$	Probability that risk factors A and B are in the states α and β respectively given that Γ is in state γ. Note: γ is an incidence state; the states labelled α and β are prevalence states
$\boldsymbol{\eta_{AB}}$	Correlation between risk factors A and B
$\boldsymbol{\eta_{AB\|\Gamma}}$	Correlation between risk factors A and B given disease Γ is incident
$\boldsymbol{\rho_{\Gamma\|_{AB}}(\gamma\|\alpha,\beta)}$	Relative risk of disease Γ given risk factor states α and β

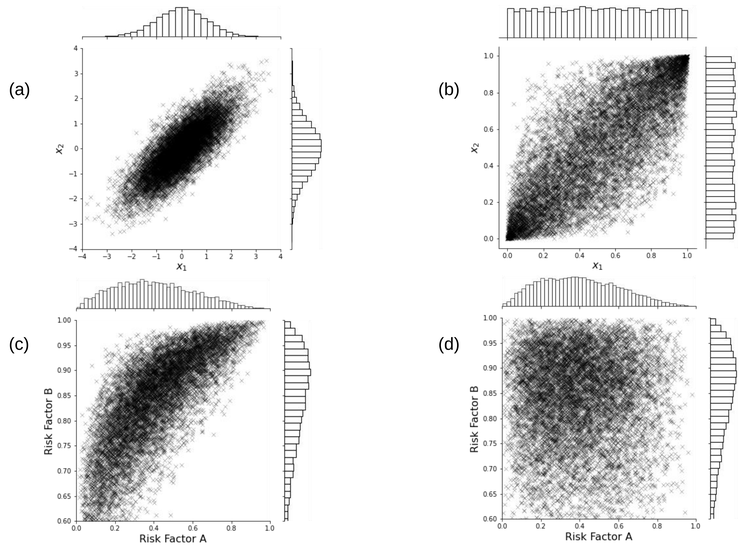
Descriptive explanation of the model

To demonstrate how the copula method works, the joint prevalence between two risk factors will be calculated.

Let: $\eta_{AB} = 0.8, \pi_A \sim Beta(2,3), \pi_B \sim Beta(10,2)$

The first step is to sample from a multivariate Gaussian with correlation set to η_{AB} = 0.8. In this case, 10,000 Monte-Carlo samples were used. Having transformed these samples using the Gaussian cumulative distributive function (CDF), the resulting distribution still has a correlation of 0.8 but now has uniform marginals. The respective inverse CDFs of each Beta distribution are then applied to one of the two dimensions of the samples to generate the required marginal distributions. Note: in public health applications, the marginal distributions may be theoretical parameterised distributions such as the Beta but are often empirical if survey data has been used. This gives the required joint prevalence distribution. If a categorical distribution was desired, the number of Monte Carlo samples between categorical bounds (e.g. Risk Factor A between values 0.2 and 0.4, Risk Factor B values between 0.7 and 0.8) can be counted and then normalised using the total number of samples (Monte Carlo integration). Also shown in **Figure 2** is the distribution that would be generated using the simple independence assumption of multiplying the one-dimensional prevalence distributions. This is equivalent to applying the marginal inverse CDFs to a uniform copula, rather than the Gaussian with correlation. Incorporation of correlation information via the copula method allows for a much more nuanced distribution to be produced.

Figure 2. Example implementation of the copula method. (a) Sample from a multivariate Gaussian with correlation ηAB (the correlation between the two risk factors). (b) The sample is transformed using the Gaussian cumulative distribution function (CFD) to generate a sample with uniform marginals while retaining the correlation – this distribution is the copula. (c) Transform this sample with the univariate inverse CDFs for each risk factor to generate the estimated joint distribution. This process ensures the approximate distribution has the appropriate correlation and marginals. The whole process is repeated for the disease population prevalence and correlation to generate a joint prevalence distribution for the disease population. (d) For comparison, shown here is the joint



Evaluation of the Copula method using the NHANES dataset

To demonstrate the use of the copula method for disease modelling purposes, joint relative risks for CLD and CHD were calculated. For both diseases, the risk factors studied were BMI and alcohol consumption. Single risk factor prevalence distributions and correlations were calculated using data from the population-level, cross-sectional, National Health and Nutrition Examination Survey (NHANES) and the copula method was used to combine them into joint prevalence. We made several assumptions while cleaning the NHANES data. Firstly, we combined a number of years of survey data into a single dataset (justified by the fact that each survey is cross-sectional so the subject data are independent). To derive an estimate for weekly alcohol consumption, we assumed a "drink" contains 8 grams of alcohol.

Estimates of the joint relative risks for CLD and CHD generated via the copula method can be found in **Table 2** and **Table 3** respectively. Joint relative risks were generally in line with acknowledged trends: [2-4]. For example, Hart et al. [4] calculated joint relative risks from longitudinal data using the same categorical definitions as those used here. However, a direct quantitative comparison is unavailable as they reported relative risks for CLD mortality whereas the results reported here are for disease incidence. Notwithstanding this, the same patterns in the relative risks are observed with relative risk seen to increase as BMI and alcohol consumption increases.

Table 2. Joint risk ratios for CLD using NHANES data

		BMI		
		Healthy Weight	Overweight	Obesie
ALCOHOL CONSUMPTION	Low risk	1.000	0.808	3.532
	Increasing risk	1.354	1.133	4.540
	High Risk	1.608	1.313	4.818

Table 3. Joint risk ratios for CHD using NHANES data

		BMI		
		Healthy Weight	Overweight	Obese
ALCOHOL CONSUMPTION	Low risk	1.000	0.490	2.322
	Increasing risk	1.470	0.731	3.580
	High Risk	1.694	0.875	4.430

Conclusion

The copula method has the potential to improve the quality of predictive models for NCDs by facilitating the inclusion of interaction effects between risk factors, without the need for expensive longitudinal studies. Although the quality of the approximation can vary depending on the input data, there are clear benefits for public health modelling. There are a number of areas for further investigation including testing on more diseases and risk factors, as well as applications within microsimulation both in the context of NCDs and other domain areas.

References: [1] : Smith M, Min A, Almeida C, et al. Modeling Longitudinal Data Using a Pair-Copula Decomposition of Serial Dependence. J Am Stat Assoc 2010; 105: 1467–1479. ; [2] Naveau S, Giraud V, Borotto E, et al. Excess weight risk factor for alcoholic liver disease. Hepatology 1997; 25: 108–111. [3] Gramenzi A, Caputo F, Biselli M, et al. alcoholic liver disease--pathophysiological aspects and risk factors. Aliment Pharmacol \& Ther 2006; 24: 1151–1161.[4]Hart CL, Morrison DS, Batty GD, et al. Effect of body mass index and alcohol consumption on liver disease: analysis of data from two prospective cohort studies. BMJ; 340.

1. HealthLumen, London, United Kingdom; 2. Liver Unit, Department of Internal Medicine, Hospital Universitario Vall d'Hebron, Barcelona, Spain; 3. Clínica Universitária de Gastrenterologia, Faculdade de Medicina, Universidade de Lisboa, Portugal; 4. Department of Hepatology and Gastroenterology , Aarhus, Denmark; 5. Barcelona Institute for Global Health (ISGlobal), Hospital Clínic, University of Barcelona, Spain; 6. Department of Pathology and Immunology of the University Hospitals of Geneva, Geneva, Switzerland; 7. Université Paris, Hospital Jean-Verdier, Paris, France; 8. Population Hepatology Research Group, University of Southampton, Southampton, UK; 9. Department of Gastroenterology, HPB Surgery and Transplantology, Military Medical Academy, Sofia, Bulgaria; 10. School of Public Health, Department of Nutrition, Health and Behavior, University of Haifa, Haifa, Israel