Improving colorectal cancer screening detection programs through Bayesian Networks and graphical modelling

Ramon Clèries[1], *Joan C. Oliva*[2] *and the Colontif Group*[3]

[1]r.cleries@iconcologia.net, Catalan Institute of Oncology – Barcelona – Spain

[2] joanc.oliva@gmail.com, Fundació Parc Taulí-Sabadell-Spain

[3]Members of the Colontif Group are: Jaume Boadas from Consorci Sanitari de Terrasa; Sara Galter from consorci Hospitalari de Terrasa; Rafel Campo Fernández de los Ríos and Eva Martínez from Fundació Parc Taulí; Victoria Gonzalo from Hospital Universitari Mutua de Terrasa; Josepa Ribes and Xavi Sanz from Catalan Plan for Oncology; Antoni Alsius from CATLAB, Viladecaballs; Fernando Fernández-Bañares from Fundació Mútua de Terrasa;

Fast-track programs to detect patients with colorectal cancer (CRC) based on high risk symptoms usually lack of sentivity and specificity. Derivation of a proper predictive score for advanced colonic neoplasm in patients with symptoms might improve fast-track CRC programs and advance diagnosis. To derive this score, a probabilistic approach based on graphical modelling has been used which allowed to identify how variables are linked and how probabilities of relationship among them can be derived. To perform this analysis and graphical representation we have used a Bayesian network (BN) analysis. This is a form of statistical modeling that stems from empirical data and by creating a graphical network that describes the best-fit dependency structure between observed variables. The key distinction between standard multivariable regression analyses and BN-type analyses is that multivariable regressions seek to identify covariates associated with a certain outcome variable whereas BN go beyond and attempt to empirically separate the associated covariates into those directly or indirectly dependent upon the outcome variable. In this way Bayesian network analyses are superior to standard approaches for inferring statistical dependencies from complex observational data. More importantly, the Bayesian best-fit dependency network structure can be used as a template for creating simulated datasets with large sample size for further posterior probability estimations for variables directly or indirectly dependent upon the outcome variable. A BN was fitted using data from Hospital centers in Barcelona area. With this data, a cohort of N=10,000 simulated patients was generated in order to provide probabilities of disease and symptoms as well as to predict the expected profiles of the CRC patients. Learning structure of data with log-linear, graphical and decomposable models for contingency tables will be presented.

Keywords: Bayesian Networks, Simulation, graphical models.

AMS: AMS Classification (Optional).