

# **Unraveling Biophysical Interactions of Radiation Pneumonitis in Non-Small-Cell Lung Cancer via Bayesian Network Analysis**

Yi Luo, Issam El Naqa, Daniel L. McShan, Dipankar Ray, Ines Lohse, Martha M. Matuszak, Dawn Owen, Shruti Jolly, Theodore S. Lawrence, Feng-Ming Kong, Randall K. Ten Haken

Affiliations of authors: Department of Radiation Oncology, The University of Michigan, Ann Arbor, MI (YL, IEN, DM, DR, IL, MS, MM, DO, SJ, TL, RTH); Department of Radiation Oncology, Indiana University, Indianapolis, IN (FMK)

Correspondence to: Randall K. Ten Haken, PhD, University of Michigan, Department of Radiation Oncology, UH B2C432, SPC 5010, 1500 East Medical Center Dr., Ann Arbor, MI 48109-5010 (e-mail: [rth@med.umich.edu](mailto:rth@med.umich.edu); phone: (734) 936-8695; Fax (734) 936-7859)

## ***Appendix B: Input Data Pre-Processing***

Along with common dosimetric information, such as Mean\_Lung\_Dose, Mean\_Heart\_Dose, and volumes receiving at least 5 Gy (V5), or 13 Gy (V13), or 20 Gy (V20) extracted from EQD2-corrected dose distributions, each patient in the dataset had a total 200 features (Table 1) from five categories including clinical factors (e.g., age, KPS, smoking status), pre-treatment cytokines (e.g., pre\_IL\_15), the slopes of cytokine changes during the treatment course (e.g., SLP\_TGF\_beta1), microRNAs (e.g., miR\_191\_5p), and SNPs (e.g., tgf\_beta1\_Rs1800469). The Hartemink's pairwise mutual information method was employed to discretize continuous variables (such as Mean\_Lung\_Dose, age, GTV) into three categories [1]; interval discretization was used for the categorical variables such as RP2 and SNPs. SNPs were described by three kinds of genotypes: wild type homozygote, minor allele homozygote, and heterozygote. In general, the radiation pneumonitis associated risk alleles were rare alleles of ancestral or derived (mutant) backgrounds. The risk allele of a SNP to cause RP2 is identified from trend tests. However, if the trend was not significant, a likelihood ratio test was conducted on only one degree of freedom to find the risk allele[2]. Statistical methods were implemented in the R-software environment and the BN learning was done using its "bnlearn" package [3].

### References

1. Hartemink AJ. Principled Computational Methods for the Validation and Discovery of Genetic Regulatory Networks. MASSACHUSETTS INSTITUTE OF TECHNOLOGY Department of Electrical Engineering and Computer Science 2001.
2. Johnson N, Fletcher O, Palles C, *et al.* Counting potentially functional variants in BRCA1, BRCA2 and ATM predicts breast cancer susceptibility. *Hum Mol Genet* 2007;16(9):1051-7.
3. Scutari W. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software* 2010;35(3).