Predication-based Bayesian network analysis of gene sets and knowledge-based SNP abstractions

Skanda Koppula

Second Annual MIT PRIMES Conference
May 20th, 2012

Mentors: Dr. Gil Alterovitz and Dr. Amin Zollanvari

Outline of Presentation

1. Introduction
    - A Predictive Model for Disease Diagnosis
    - Alcohol Dependency and Lung Cancer
2. Methodology
    - An Overview of Prediction Based Analysis
    - SNP to Gene Mapping
    - Gene Set and Training Data + Parallelization
3. Results
    - AUROCS for Alcohol Dependence
    - AUROCS for Lung Cancer
4. Discussion
5. Future Work

A Predictive Model of Disease Diagnosis

Top-Level Goals *for Genome Wide Association Studies (GWAS)*

Understand underlying mechanisms behind disease

Infer diagnosis from patient's biological data

A Predictive Model of Disease Diagnosis

Top-Level Goals *for Genome Wide Association Studies (GWAS)*

Understand underlying mechanisms behind disease

Infer diagnosis from patient's biological data

Methods

Gene Set Enrichment Analysis [1]

Finds genes in set S (~cellular pathway) in top/bottom of ranked list of genes – ordered by importance in classifying

Predictive-Based Gene Set Analysis

Finds predictive accuracy of S via probabilistic relations between of S to disease (Bayesian Network)

Is this accuracy is *significant* compared to random prediction?

If so, network can be used in disease diagnosis.

Alcohol Dependence and Lung Cancer

Goal

Use predictive-based analysis for both gene and SNP expression data

Analyze alcohol dependency and lung cancer

Alcohol Dependence and Lung Cancer

Goal

Use predictive-based analysis for both gene and SNP expression data
Analyze alcohol dependency and lung cancer

Alcohol Dependence (Alcoholism)

- Underlying biological pathways not identified
- Difficult to overcome once dependence is initiated:
- Affects 140 million people

Lung Cancer

- Other diagnostic methods may be invasive
  - Early accurate diagnosis improves chances of survival
- Causes of death for 160,000 people per year

Alcohol Dependence and Lung Cancer

Goal

Use predictive-based analysis for both gene and SNP expression data

Analyze alcohol dependency and lung cancer

Alcohol Dependence (Alcoholism)

- Underlying biological pathways not identified
- Difficult to overcome once dependence is initiated:
- Affects 140 million people

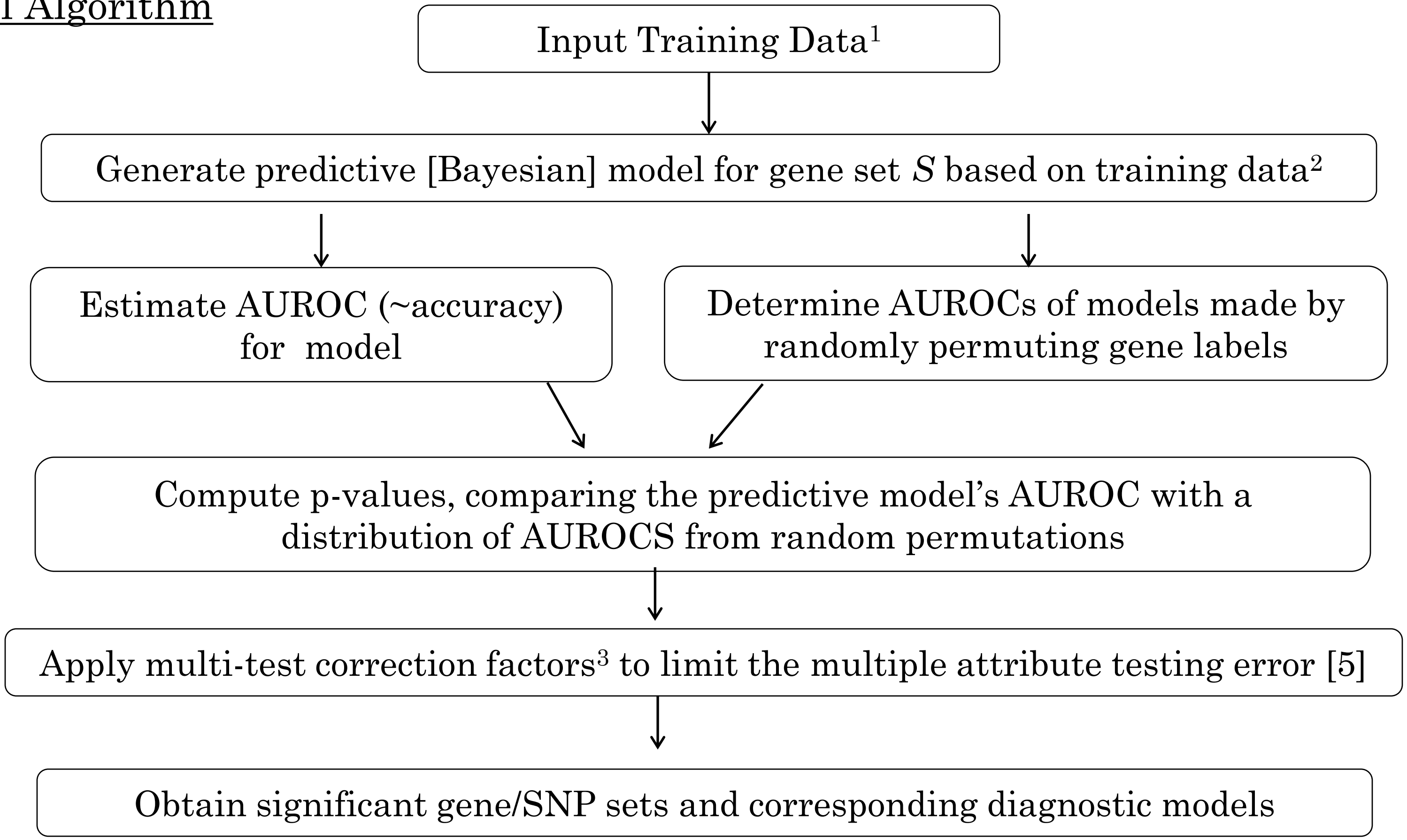
Lung Cancer

- Other diagnostic methods may be invasive
  - Early accurate diagnosis improves chances of survival
- Causes of death for 160,000 people per year

Focus on *robustness* and accuracy – same sets identified as significant, across independently collected clinical data

An Overview of Prediction Based Analysis

General Algorithm



Input Training Data[1]

Generate predictive [Bayesian] model for gene set *S* based on training data[2]

Estimate AUROC (~accuracy) for model

Determine AUROCs of models made by randomly permuting gene labels

Compute p-values, comparing the predictive model's AUROC with a distribution of AUROCS from random permutations

Apply multi-test correction factors[3] to limit the multiple attribute testing error [5]

Obtain significant gene/SNP sets and corresponding diagnostic models

[1] – Patients' biological data. Ex. the set of gene expression values for each patient

[2] – Network creation from training data done via machine learning tool WEKA

[3] – Applied corrections: Benjamini-Hochberg FDR, Bonferroni, and Storey FDR

Implementation of algorithm and WEKA in Java

A SNP to Gene Mapping

Can we analyze SNP data to create diagnostic models?

- SNP is a single nucleotide polymorphism: two character DNA mutation
- We can determine specific SNP profiles that imply disease

Yes: consider SNP sets to be analogous to gene sets and apply previous algorithm implementation

A SNP to Gene Mapping

Can we analyze SNP data to create diagnostic models?

SNP is a single nucleotide polymorphism: two character DNA mutation

We can determine specific SNP profiles that imply disease

Yes: consider SNP sets to be analogous to gene sets and apply previous algorithm implementation

To create such sets, we can develop a 1:1 map from a SNP to a particular gene

The mapping pairs each SNP to its closest gene

(character distance in DNA string sequence)

A SNP to Gene Mapping

Can we analyze SNP data to create diagnostic models?

SNP is a single nucleotide polymorphism: two character DNA mutation

We can determine specific SNP profiles that imply disease

Yes: consider SNP sets to be analogous to gene sets and apply previous algorithm implementation

To create such sets, we can develop a 1:1 map from a SNP to a particular gene

The mapping pairs each SNP to its closest gene

(character distance in DNA string sequence)

Possible biological meaning of SNP to gene mapping: the value of the SNP may affect the function and expression of the gene closest to it.

Relevance of being able to analyze SNP data:

- Only data for a disease may be SNP data
- Insight on the biological significance of SNPs

Gene Set and Training Data + Parallelization

We used three sets the source of the tested gene-sets

from the KEGG, GO, and the curated set used by GSEA's creators (as a comparison)

Alcoholism training data: COGA - Collaborative Study on Genetics of Alcoholism
COGEND - Collaborative Genetic Study of Nicotine Dependence

Lung cancer training data: Boston study - National Medicine Labs
Michigan study - National Medicine Labs

Gene Set and Training Data + Parallelization

We used three sets the source of the tested gene-sets

from the KEGG, GO, and the curated set used by GSEA's creators (as a comparison)

Alcoholism training data: COGA - Collaborative Study on Genetics of Alcoholism

COGEND - Collaborative Genetic Study of Nicotine Dependence

Lung cancer training data: Boston study - National Medicine Labs

Michigan study - National Medicine Labs

In COGA and COGEND:

Each patient had values for each of one million SNPs

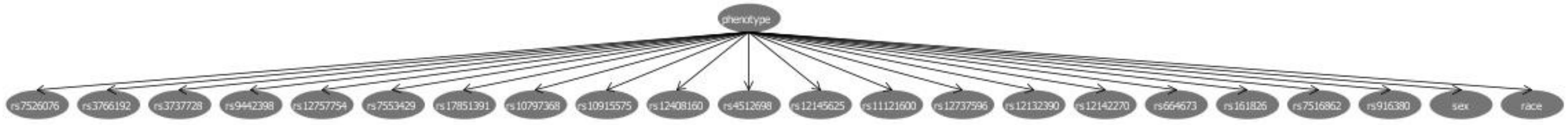
Total of 3,600 patients

To avoid having to reduce data (time and computing limitations), we parallelized the creation of the SNP-to-gene mapping and partially parallelized segments of the prediction based algorithm implementation

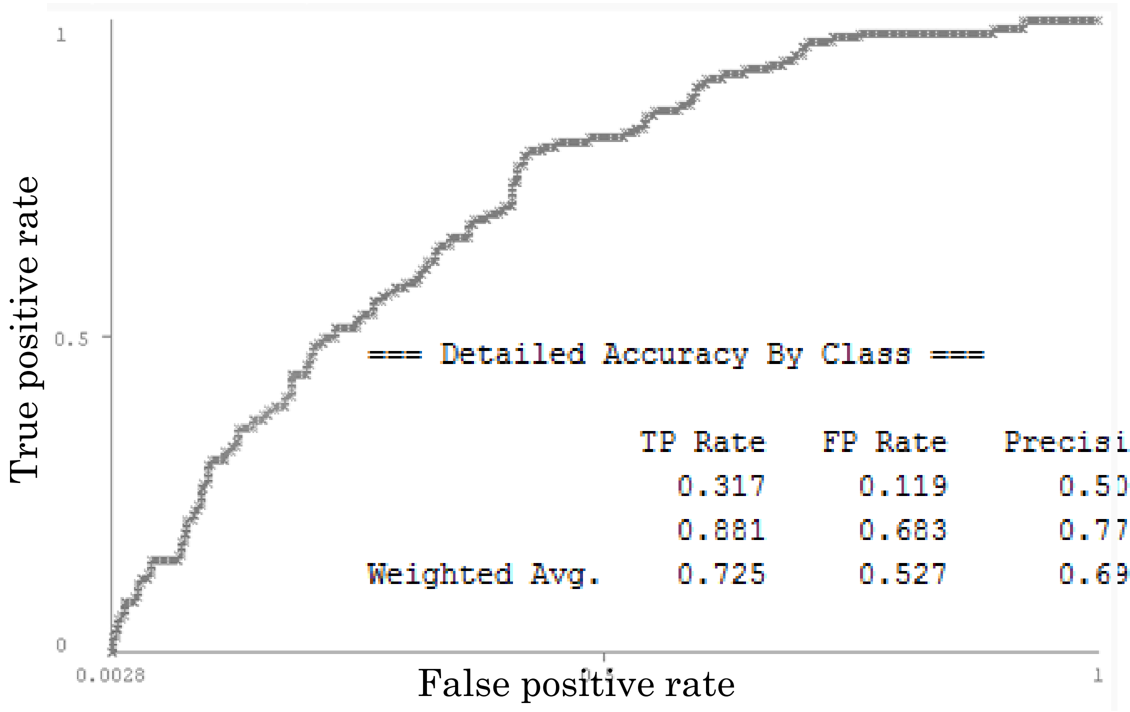
Results: Alcohol Dependency – Example

NaiveBayesian network for one identified significant SNP set:

BIOGENIC_AMINE_METABOLIC_PROCESS



Area Under Receive-Operating-Curve ~ Accuracy ~ 70.2%



```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.317     0.119     0.506      0.317     0.389      0.716    1N
                 0.881     0.683     0.771      0.881     0.822      0.716    2N
Weighted Avg.    0.725     0.527     0.697      0.725     0.702      0.716
```

Results: Alcohol Dependence

A total of 15 significant *shared* gene/SNP sets were found

(significant using both the COGA and COGEND training sets)

Gene Set Name	Mean AUROC	Mean Significance Value	
BRAIN_DEVELOPMENT	0.651562	0.011	GO Gene Set Repository
IMMUNE_EFFECTOR_PROCESS	0.654478	0.022	
INTERFERON_GAMMA_BIOSYNTHETIC_PROCESS	0.666999	0.033	
DEFENSE_RESPONSE_TO_VIRUS	0.664685	0.016	
INTERFERON_GAMMA_PRODUCTION	0.667297	0.024	
AMINO_ACID_DERIVATIVE_METABOLIC_PROCESS	0.661525	0.011	
BIOGENIC_AMINE_METABOLIC_PROCESS	0.702950	0.008	
REGULATION_OF_INTERFERON_GAMMA_BIOSYNTHETIC_PROCESS	0.666999	0.033	
ALCOHOL_METABOLIC_PROCESS	0.657533	0.001	
CENTRAL_NERVOUS_SYSTEM_DEVELOPMENT	0.651999	0.016	
INORGANIC_ANION_TRANSPORT	0.658397	0.005	
POSITIVE_REGULATION_OF_CYTOKINE_BIOSYNTHETIC_PROCESS	0.666273	0.019	
PEPTIDE_METABOLIC_PROCESS	0.660818	0.007	
KEGG_DILATED_CARDIOMYOPATHY	0.667531	0.014	KEGG
KEGG_ARGININE_AND_PROLINE_METABOLISM	0.659727	0.001	
KEGG_VIRAL_MYOCARDITIS	0.718438	0.001	
KEGG_PROXIMAL_TUBULE_BICARBONATE_RECLAMATION	0.654189	0.029	

Results: Alcohol Dependence

Some are biologically interesting…

Gene Set Name	Note
BRAIN_DEVELOPMENT	Association confirmed from *in-vivo* by Maier *et al.* for fetal[6]
IMMUNE_EFFECTOR_PROCESS	Kronfol et al.[7]
INTERFERON_GAMMA_BIOSYNTHETIC_PROCESS	Jeong et al. [8]
KEGG_VIRAL_MYOCARDITIS	Wilke et al. [9]
KEGG_DILATED_CARDIOMYOPATHY	*Dilated cardiomyopathy defined to be caused by alcoholism*
KEGG_ARGININE_AND_PROLINE_METABOLISM	New association?
KEGG_PROXIMAL_TUBULE_BICARBONATE_RECLAMATION	New association?
ALCOHOL_METABOLIC_PROCESS	Associated by generality
PEPTIDE_METABOLIC_PROCESS	New association?
INORGANIC_ANION_TRANSPORT	Related to PROXIMAL_TUBULE…

Median AUROC of all 15 sets………… **0.6615**

Median COGA AUROC……………….. **0.6854**
Median COGEND AUROC……………. **0.6588**

Number of significant sets:
From COGA: 28
From COGEND: 35

Results: Lung Cancer

Some of these 15 significant common* gene/SNP sets are biologically interesting.

Gene Set Name	M. AUROC	M. Significance Value	
ACTIN_FILAMENT_BASED_MOVEMENT	0.653367	0.03	GO
G1_PHASE	0.662136	0.0265	
INTRACELLULAR_SIGNALING_CASCADE	0.657008	0.033	
DEVELOPMENTAL_MATURATION	0.640078	0.0445	
ACTIN_FILAMENT_ORGANIZATION	0.695239	0.013	
KEGG_CIRCADIAN_RHYTHM_MAMMAL	0.658049	0.03	KEGG
KEGG_GLYCOLYSIS_GLUCONEOGENESIS	0.702448	0.011	
MAP00010_Glycolysis_Gluconeogenesis	0.683253	0.0175	GSEA*
P53_DOWN	0.649584	0.041	
P53_UP	0.675317	0.0215	

Median AUROC of all 10 sets............ **0.6609**

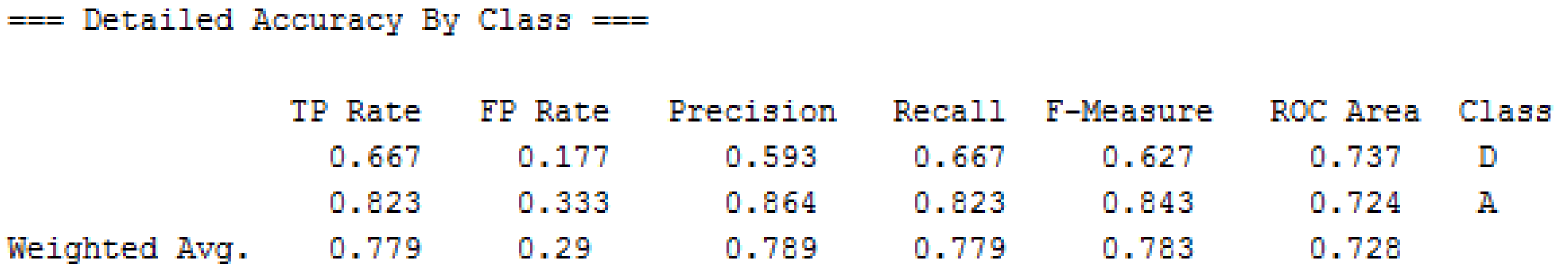
*Uses the set of gene sets used by GSEA (standard for comparison) [1]

Number significant sets:

Boston: 95
Michigan: 74

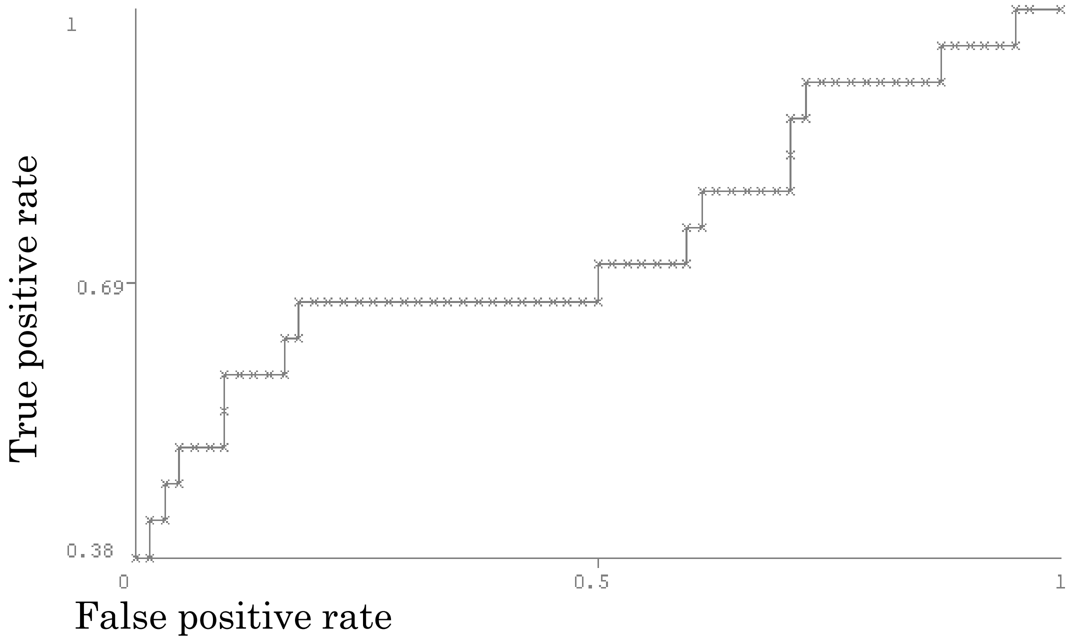
Results: Lung Cancer

If we create a network encompassing all found lung pathways…



```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.667     0.177      0.593     0.667     0.627      0.737    D
                   0.823     0.333      0.864     0.823     0.843      0.724    A
Weighted Avg.      0.779     0.29       0.789     0.779     0.783      0.728
```



Conclusions

- Moderately high AUROCs (0.6 to 0.75) and significance of sets

  → ‘good’ accuracy of diagnostic models – individual and combined [general AUROC metric]

Conclusions

- Moderately high AUROCs (0.6 to 0.75) and significance of sets

  → 'good' accuracy of diagnostic models – individual and combined [general AUROC metric]

- Identified associations confirmed. – e.g. VIRAL_MYOCARDITIS [9]
- New ones found. – e.g. ARGININE_AND_PROLINE_METABOLISM

  Alchoholism: *identified* 15 significant, robust (data-independent) pathways
  Lung Cancer: *identified* 10 significant, robust pathways

The gene/SNP sets → cellular pathways;
significance in disease prediction → role in disease's biological mechanism

Conclusions

- Moderately high AUROCs (0.6 to 0.75) and significance of sets

  → 'good' accuracy of diagnostic models – individual and combined [general AUROC metric]

- Identified associations confirmed. – e.g. VIRAL_MYOCARDITIS [9]
- New ones found. – e.g. ARGININE_AND_PROLINE_METABOLISM

  Alchoholism: *identified* 15 significant, robust (data-independent) pathways
  Lung Cancer: *identified* 10 significant, robust pathways

The gene/SNP sets → cellular pathways;
significance in disease prediction → role in disease's biological mechanism

- Good robustness for select data of predictive based analysis

  (higher number of significant pathways shared by COGA and COGEND)

- 40, 41 significant pathways vs. the <8, 11 pathways identified by GSEA [1]
  - Robustness (num. common pathways) similar in value

Future Work

- *In vivo* testing of biologically significant pathways
- Develop new gene-protein interaction and other bio. hypothesis

Future Work

- *In vivo* testing of biologically significant pathways
- Develop new gene-protein interaction and other bio. hypothesis

- Consider the results from other training data sets; further confirm and statistically quantify method robustness
- Model other diseases

Future Work

- *In vivo* testing of biologically significant pathways
- Develop new gene-protein interaction and other bio. hypothesis

- Consider the results from other training data sets; further confirm and statistically quantify method robustness
- Model other diseases

- Optimize the mechanics and implementation of prediction based algorithm
- Factor in the importance of cliques within the networks
  - Establish gene-interdependency - networks such as Augmented NaïveBayes

Acknowledgements

Thank you:

Dr. Gil Alterovitz for his insightful guidance on the project's direction and challenges

Dr. Amin Zollanvari for his great mentorship - guiding me to discover answers to my many questions and helping me when I had difficulties

Dr. Tanya Khovanova for her useful project guidance and advice

My parents for always being supportive

References

[1] Gene set enrichment analysis: A knowledge based approach for interpreting genome-wide expression profiles. *Subramanian et al.* 2005.

[2] Prediction-based Bayesian Network Analysis of Gene Sets for Genome-wide Association and Expression Studies. *Zollanvari and Alterovitz*. 2012.

[3] Letter from the WHO European Ministerial Conference on Young People and Alcohol. *Brundtland and World Health Organization*. 2001.

[4] Lung cancer (small cell) overview. *American Cancer Society*. 2012.

[5] Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Benjamini and Hochberg*. 1995.

[6] Fetal Alcohol Exposure and Temporal Vulnerability: Regional Differences in Cell Loss as a Function of the Timing of Binge-Like Alcohol Exposure During Brain Development. *Maier et al.* 1999.

[7] Immune Function in Alcoholism: A Controlled Study. *Kronfol et al.* 1993.

[8] Abrogation of the antifibrotic effects of natural killer cells/interferon-gamma contributes to alcohol acceleration of liver fibrosis. *Jeong et al.* 2008.

[9] [Alcohol and myocarditis] *Wilke et al.* 1996.

Thank you for your attention!