Graphical Modeling Driven Policy Insights for Household Air Pollution: Insights from 204,912 Patients' POSEIDON Study

Tavpritesh Sethi* TAVPRITESHSETHI@GMAIL.COM
All India Institute of Medical Sciences, Ansari Nagar, New Delhi
Anurag Agrawal A.AGRAWAL@IGIB.IN
Aditya Nagori ADITYA.KUMAR@IGIB.IN
CSIR-Institute of Genomics and Integrative Biology, Mall Road, Delhi
Suveen Angraal SUVEEN.AIIMS@GMAIL.COM
All India Institute of Medical Sciences, Ansari Nagar, New Delhi
Sundeep Salvi SSALVI@CRFINDIA.COM
Chest Research Foundation, Pune

Abstract

While healthcare researchers intuitively recognize the connectivity of physiological systems, data-driven decision models for such interactions are lacking. Graphical models are amongst the best techniques of capturing interactions in complex interconnected data. In this work we leveraged the strengths of probabilistic graphical models applied to a massive, one-of-its-kind clinical data of 204,912 patients collected all across India on a single day to derive policy insights into Household Air Pollution associated respiratory disease. This unique data resource, called POSEIDON (Prevalence Of Symptoms on a singlE Indian healthcare Day On a Nationwide scale) documented 60 clinical symptoms and diseases in 204,912 patients visiting the OPDs of 7400 randomly sampled general practitioners on a single day, all across India. We constructed "Symptome maps" across the age-groups of patients using Association Networks, discovered dynamic community structures and visualized these as Alluvial maps. These revealed novel qualitative insights into comorbidity patterns of various diseases and confirmed the high comorbid occurrence of airway diseases across all age groups. To quantify the effect of unclean cooking fuel upon airway disease, we fused these data with district wise Liquefied Petroleum Gas (LPG) consumption and used Bayesian structure learning and inference algorithms to quantify the causal impact of clean cooking fuels (LPG) on lowering Obstructive Airway Disease (OAD). Further application of approximate inference upon the nodes of interest showed that the rate of OAD could be lowered by as much as 50% if the LPG usage is increased from the lowest to the highest quartile. To our knowledge, this is the first of its kind application of graphical models to a practice based clinical data in India combined with an indicator of clean energy usage to yield policy insights.

1. Introduction

The WHO UNDP 2016 report (http://www.who.int/indoorair/en/) lists Household Air Pollution (HAP) as a global emergency and the single most important global environmental health problem. HAP causes 4.3 million premature deaths each year over the globe. India has the largest absolute contribution to these premature deaths (1.3 million premature deaths every year) with almost 800 million people depending upon unclean cooking fuel. HAP causes morbidity right from the intra-uterine life by causing growth retardation and has been shown to be associated with still-births, perinatal mortality, asthma, respiratory infections including tuberculosis, cardiac disorders, cataract and cancers of the mouth and nasopharynx. In addition, contrary to what is thought, HAP is not only responsible for indoor air pollution but contributes as high as 30% to Outdoor Air Pollution as well. Therefore, HAP is problem of immediate concern that needs to be addressed. Despite various efforts, the quantification of disease burden attributable to unclean cooking fuel has remained difficult due to the complex interplay of factors. We leveraged the strengths of association graphs and probabilistic graphical modeling applied to a massive, one-of-its-kind clinical data of 204,912 patients collected all across India on a

Proceedings of the 2nd Indian Workshop on Machine Learning, IIT Kanpur, India, 2016.

single day to derive policy insights into the respiratory morbidity. This unique data resource, called the POSEIDON study (Prevalence Of Symptoms on a singlE Indian Healthcare Day On a Nationwide Scale (Salvi et al., 2015) documented 60 clinical symptoms and diseases in 204,912 patients visiting the Out Patient Departments (OPDs) of 7400 General Practitioners randomly selected across India. We hypothesized that symptoms and diseases will form community structures (*Symptome*) which may vary across age groups of Indian patient base. Further, we tested the utility of fusing POSEIDON data with energy indicators such as consumption of Liquefied Petroleum Gas (LPG) to yield quantitative insights into clean cooking fuel and its relationship with the "Symptome". We show that these insights could be important in driving the HAP policy in India.

2. Methods

2.1 Association Networks and Community Detection

POSEIDON data were programmatically inspected and cleaned for anomalies using R programming language (R (Foundation for Statistical Computing, Vienna, Austria). Data were then stratified into decades and pair-wise associations of symptoms were computed in R using Fisher's exact. The negative logarithms of the p-values obtained (for all $p<0{\cdot}05$) were used as weights in the weighted edgelist and community detection was done using Infomap algorithm (Rosvall et al., 2011). Finally, the age-wise modular changes were visualized as an alluvial diagram using alluvial generator (MapEquation, Umeå, Sweden)

2.2 Data Driven Structure Learning Using Bayesian Networks

To quantify the role of clean cooking fuel on HAP, we merged the clinical information in POSEIDON with the data on district-wise LPG consumption in India available from the United Nations Statistics Division, DevInfo. The data were discretized using a quantiles. Structure learning was performed using Tabu search (Glover, 1989) which is a score based structure-learning algorithms implemented in R using bnlearn package (Scutari, 2010).

2.3 Bayesian Network Inference

Approximate inference method was used to determine the conditional probabilities in the joint multivariate network model. Conditional inference was carried out on LPG usage and the rates of Obstructive Airway Disease in the district wise POSEIDON data

3. Results and Discussion

3.1 Association Networks and Community Structure Within the Symptome

The network and alluvial graph generated upon the POSEIDON data captured the relationships between different disease condition and symptoms. These statistically significant associations, however were found to be dynamically changing across the age age groups (Figure 1.a). Further, these associations also led to the formation of community structures within the "Symptome" (Figure 1.a) and their evolution was visualized in the as an alluvial diagram (Figure 1.b). Respiratory comorbidities were the most predominant reason for visiting a GP across all age groups in India, the visualization showed the heaviest contribution from the respiratory module (blue streamline). In addition to respiratory disease which forms the further focus of this analysis, this visualization also confirmed other medical rhetoric such as differential co-evolution of heart-disease with diabetes and anemia in the reproductive age group in females (which is hugely prevalent in India). This is seen as the merging of bands labeled *Female Genital* (violet) and *Anemia* (light-blue) in the reproductive age group and their separation thereafter. Similar connections were seen in the alluvial graph between male genital symptoms and urology in patients who were 50 years or older.

3.2 Bayesian Network Structure Learning and Inference

The association network modeling provided qualitative insights which, however, do not guarantee causal relationships. This motivated the use of more quantitative and causal modeling tools such as Probabilistic Graphical Models (PGMs). We fused the POSEIDON data with district-wise LPG consumption data available from United Nations Statistics Division (DevInfo). As the number of samples were high enough, we performed ab-initio Bayesian network structure learning through an unsupervised analysis. Bayesian networks being joint multivariate models, these minimize the false discovery rate that plagues multiple hypothesis testing in association networks. Tabu search was used to prevent trapping into local minima. Interestingly, the structure learnt showed a direct parent-child relationship between LPG consumption and Obstructive Airway Disease (OAD) as shown in Figure 2. Approximate inference on this node revealed that increasing the LPG penetrance from the lowest to the highest quartile in the country can lower the rate of Obstructive Airway Disease by as much as 50%. The wide range (2-85 units) of per-capita LPG

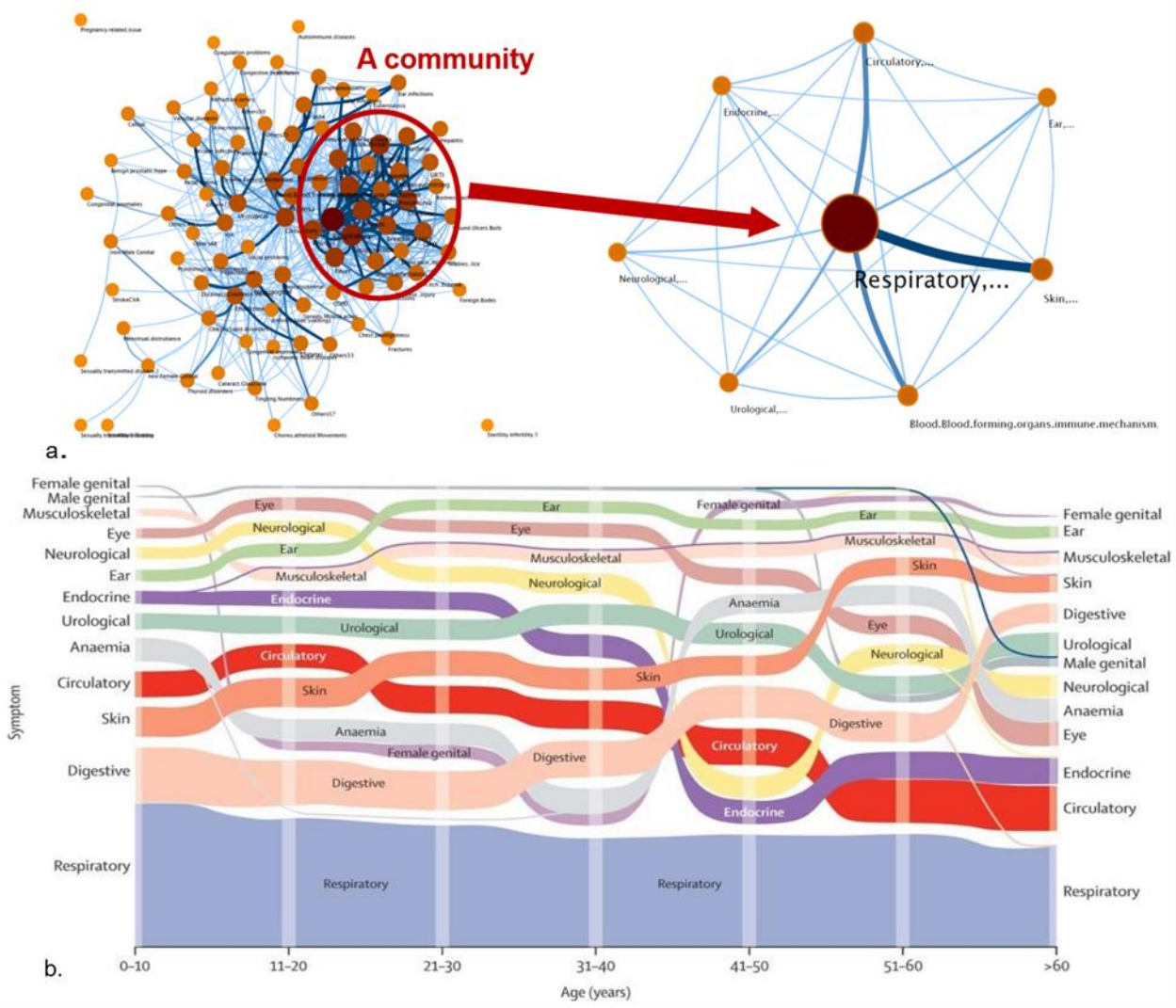

Figure 1. Association Networks analysis and visualization to gain qualitative insights into health-care policy. a) Comorbid diseases (circles) merge together (edges, blue) and form "friendship groups" called communities (right). b) These communities can be easily seen to change across the age groups of the population in the visualization in the form of an alluvial mapping. It is noteworthy that the respiratory community of diseases was found to be the most prevalent (thick blue streamline) all across the age groups of 2,04,912 Indian OPD patients.

consumption was reflective of the actionable gap to prevent this irreversible disease. A rational policy design has the quality of being holistic yet focused. With societal systems becoming increasingly complex at a rapid pace, expert based decisions have become challenging. Therefore, classical statistical now need to be supported by methods that leverage computational and mathematical advances in data science. Practice-based morbidity surveys such as the POSEIDON study provide a rich resource to discover actionable insights. This paper illustrates the key role of visualization coupled with joint inference to enable organization of complex information into interpretable patterns, hence facilitating decision making and policy.

4. Dual Submissions

Association networks analysis of POSEIDON has been published earlier in Dec 2015 issue of Lancet Global Health; Bayesian network analysis of POSEIDON is not published elsewhere or currently under review.

Acknowledgments

Authors acknowledge the funding and support provided by the Wellcome Trust/DBT India Alliance and CSIR, India. We also acknowledge Chest Research Foundation, Pune for generating and sharing POSEIDON data.

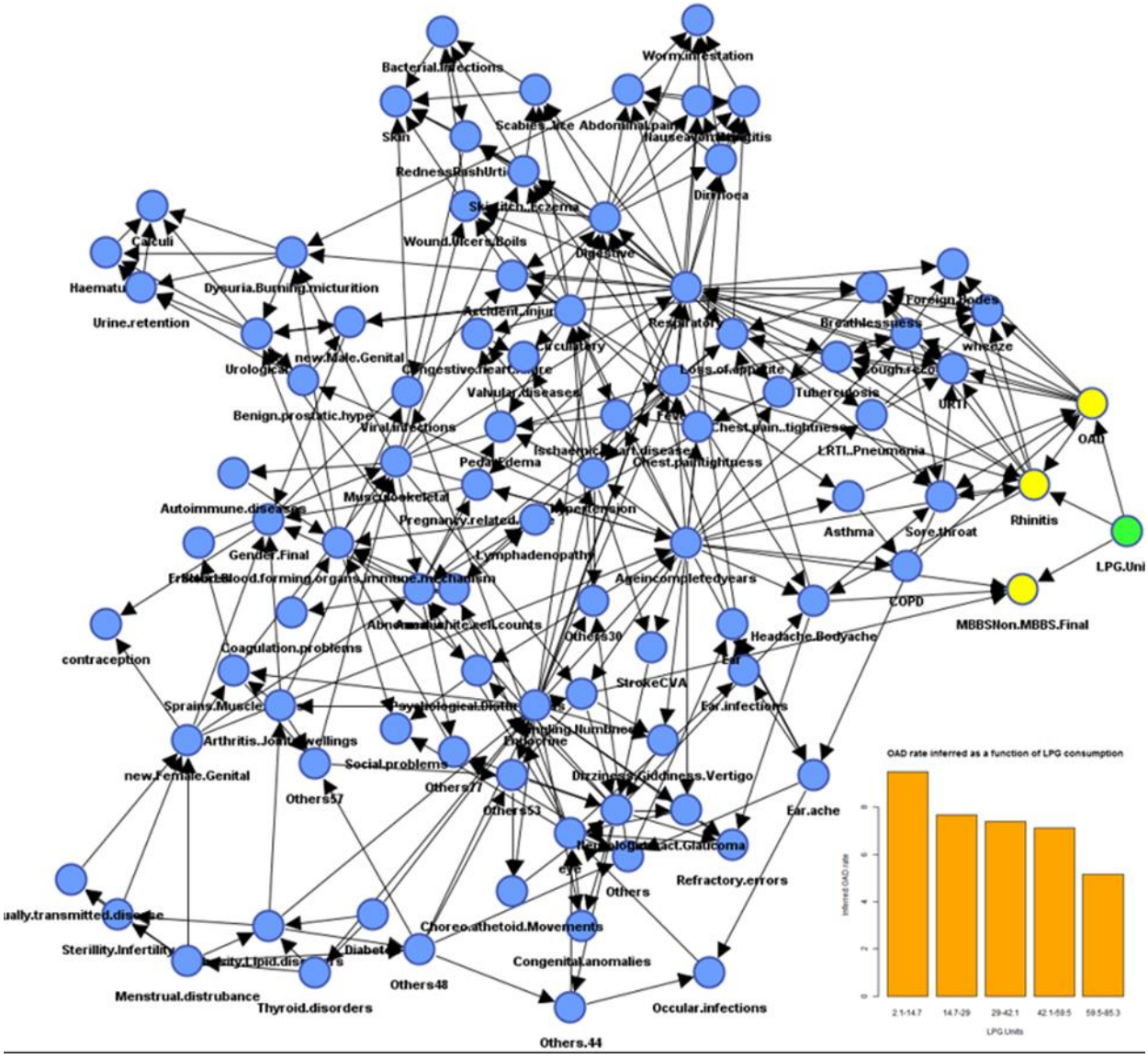

Figure 2. Bayesian Network Analysis discovered the causal influence of LPG consumption on Obstructive Airway Disease. The Bayesian Network is an AI algorithm which helps discovers the directions (arrows) of influence from the data itself. Interestingly, the LPG node is a parent (causal influence) of Obstructive Airway Disease as discovered in an entirely ab-initio fashion. The barplot (inset) shows that the rate of inferred OAD almost falls by 50% if LPG penetration is increased from the lowest bin to the highest bin across India

References

Glover, F. (1989). Tabu Search—Part I. *ORSA Journal on Computing, 1(3)*, 190-206. doi:10.1287/ijoc.1.3.190

Rosvall, M., & Bergstrom, C. T. (2011). Multilevel Compression of Random Walks on Networks Reveals Hierarchical Organization in Large Integrated Systems. *PLoS ONE, 6*(4). doi:10.1371/journal.pone.0018209

Salvi, S., Apte, K., Madas, S., Barne, M., Chhowala, S., Sethi, T., . . . Gogtay, J. (2015). Symptoms and medical conditions in 204 912 patients visiting primary health-care practitioners in India: A 1-day point prevalence study (the POSEIDON study). *The Lancet Global Health, 3*(12). doi:10.1016/s2214-109x(15)00152-7

Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software J. Stat. Soft., 35*(3). doi:10.18637/jss.v035.i03