Operationalizing an Ecological Model of College Health: Some Research Opportunities and Challenges in Informatics, Data Analytics and Machine Learning

Dr. Vasant G. Honavar
Professor of Information Sciences and Technology
Director, Center for Big Data Analytics and Discovery Informatics
301A Info Sci. and Tech. Building
Penn State University
University Park, PA 16802
vhonavar@ist.psu.edu
http://vhonavar.ist.psu.edu

Our conceptual framework for the College Student Health project is grounded in an ecological model that aims to elicit how interacting influences at the individual, social, built and natural environment, and policy levels impact health outcomes [1]. At the individual level, factors such as genetics, demographic characteristics (e.g., age, sex, race/ethnicity, socioeconomic status), behaviors (i.e., diet, smoking, drinking, exercise), have been shown to influence health outcomes. However, because individual factors explain only part of the variation in health outcomes, improving the health of college students depends upon understanding the influence of broader environmental contexts in which residents live, work, and play. Individuals are nested within a broader social environment, defined both by those with whom individuals interact with, and the specific patterns of interactions between individuals (social networks). Individuals are influenced by the demographic characteristics (e.g., neighborhood-level socioeconomic status and education) and behaviors of their peers. For instance, the prevalence of preventive health behaviors (e.g., safer sex, physical activity) among family members, friends, coworkers, and neighborhoods/communities can establish social norms that influence each individual's likelihood of adopting and sustaining such behaviors. The degree to which preventive health behaviors are socially normative can also influence the extent to which individuals will be reinforced for continuing to engage in such preventive behaviors. The social environment is nested within the broader built environment, defined by man-made influences such as the availability of parks, bike trails, and other opportunities to engage in health promoting leisurely activities, communications technology, and physical institutions, as well as the characteristics of the natural environment e.g., fresh water, forests, and air.

Our notion of built and natural environment also includes the results of interactions between environments, such as environmental contaminants that interact with air or water to produce pollution. The physical infrastructure of the built and natural environment (e.g., neighborhood walkability, access to healthy foods, air pollution) can create opportunities and limits that impact individual and community health outcomes. For instance, environmental exposures may interact with genetic predispositions to influence how genes are expressed over time (epigenetics). Similarly, access to pedestrian-friendly neighborhoods can improve health outcomes such as overweight/obesity. The model also extends prior ecological models by incorporating modern information technologies (mHealth, social media) as part of the built environment. Finally, all ecological levels are nested within a surrounding policy environment, defined both by formal policies (e.g., the Patient Protection and Affordable Care Act), and informal policies (e.g., cultural norms, or quality/performance standards). Existing research suggests that both formal policies (e.g., restrictions on workplace smoking), as well as informal policies (e.g., home smoking bans) can impact public health.

Operationalizing the the above conceptual model presents several challenges in computer science, informatics, data analytics, and machine learning:

- The underlying data are necessarily heterogeneous, high dimensional, longitudinal, and the measurements are sparse, available at varying levels of granularity and spatio-temporal resolution. The data are far from IID (independent and identically distributed), exhibit complex relational structure, and are often only partially observable. Exiting approaches to analysis of longitudinal data often make assumptions that are unlikely to hold in such settings. This calls for the formulation of

novel probabilistic models as well as efficient algorithms for learning predictive models from such data.

- Existing tools for causal discovery, e.g., causal diagrams, which provide a formal representation for combining data with causal information, and do-calculus, which provides the inferential machinery for causal inference, are currently limited in their ability to handle relational and temporal data. This calls for advances in methods for learning causal effects from relational and temporal data [2].
- Large-scale multi-site studies present the challenge of integrating the observations and experiments in disparate settings (e.g., an elite private university versus a large public university). This calls for sound statistical and computational methods and tools to support such studies [4-6].
- Last, but not the least, much of the data are needed are sensitive in nature and raise legitimate privacy concerns [6]. The lack of practical frameworks for analysis of sensitive data in a manner that does not violate applicable data access and use policies constitutes a significant barrier to the engagement of researchers with expertise in analytics in developing and evaluating advanced methods for analysis of such data, assessing the performance of alternative approaches, or ensuring the reproducibility of results. This calls for the development of computable representations of data access and use policies (DAUP) and the development of platforms and tools that support DAUP-compliant analysis of sensitive data [8].

References

1. Hovell MF, Wahlgren DR, Adams MA. The logical and empirical basis for the behavioral ecological model. In: DiClemente RJ, Crosby R, Kegler M, eds. Emerging Theories and Models in Health Promotion Practice and Research. 2nd Edition ed. San Francisco: Jossey Bass; 2009:415-449
2. Bui, N., Yen, J. and Honavar, V. (2015). Temporal Causality of Social Support in an Online Community for Cancer Survivors. In: International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (SBP15). Springer-Verlag Lecture Notes in Computer Science, Vol. 9021, pp. 13-23.
3. Bareinboim, E., Lee, S., Honavar, V. and Pearl, J. (2013). Transportability from Multiple Environments with Limited Experiments. In: Advances in Neural Information Systems (NIPS) 2013. pp. 136-144.
4. Lee, S. and Honavar, V. (2013). Transportability of a Causal Effect from Multiple Environments. In: Proceedings of the 27th Conference on Artificial Intelligence (AAAI 2013).
5. Lee, S. and Honavar, V. (2013). Causal Transportability of Experiments on Controllable Subsets of Variables: z-Transportability. In: Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI 2013).
6. Avancha, Sasikanth, Amit Baxi, and David Kotz. "Privacy in mobile technology for personal healthcare." *ACM Computing Surveys* 45.1 (2012): 3.
7. Honavar, V. (2015) EAGER: Towards a Computational Infrastructure for Analysis of Sensitive Data. http://www.nsf.gov/awardsearch/showAward?AWD_ID=1551843