Research Statement

Wenli Zhang, Ph.D. Candidate
MIS Department, Eller College of Management, University of Arizona
E-mail: wenlizhang@email.arizona.edu
Personal website: http://www.u.arizona.edu/~wenlizhang/

My research interests revolve around the areas of Data Science and Information System Design, especially in developing techniques based on Machine Learning (ML), Natural Language Processing (NLP), Network Analysis, and Distributed Computing for solving real-world problems within the context of healthcare and other business concerns. My research focus is on developing innovative information technologies to lower healthcare costs, improve healthcare process, and improve chronic disease control.

My current dissertation work focuses on the role of Big Data analytics in chronic disease management and prevention. Chronic diseases lead to lower life quality as well as increase hospitalization, long-term disability, and mortality. In the United States (US), for example, about half of all adults have at least one chronic condition; their treatments account for 86% of all US-based healthcare costs [1]. The debilitating and costly effects of chronic conditions can often be prevented or mitigated. From the perspective of health providers, attaching a high priority to patient supervision and population-level health management can significantly reduce the average costs of chronic diseases and hospital readmissions. From the perspective of patients with chronic diseases, proactive self-management could remarkably improve clinical outcomes and reduce the costs of such care. Nevertheless, researchers have depended on empirical studies or traditional survey-based methods to obtain information needed for chronic disease studies, making large scale research prohibitively expensive and slow. Over the past decades, the emergence of Big Data opens up new possibilities for promoting chronic disease control. This emerging capability has the potentials to (1) provide evidence-based medicine, (2) decrease the costs of healthcare, and (3) produce timely and customized healthcare services. However, the promise of Big Data is not without its challenges. First, the accuracy of data-driven medical decisions depends on the appropriateness of the data being used. It is often difficult to identify the right data and determine how best to use it. Second, the data are fraught with noise, and extracting the right signals from them presents technical obstacles. Finally, security and privacy concerns are additional barriers preventing researchers from taking full advantage of Big Data analytics in this context. In light of the potential of leveraging Big Data to reveal associations, patterns, and trends that would otherwise be hidden, coupled with the challenges associated with healthcare-related Big Data analytics, using Big Data for chronic disease management is now one of the fastest growing areas with interesting research challenges.

By choosing asthma (one of the most serious chronic diseases in the US) as a research case, my dissertation is to further explore the role of Big Data in healthcare analytics. I aim to investigate its applications in chronic disease surveillance, risk factor analyses and evaluating possible substitutes for certain chronic disease risk behavior by adapting advanced ML, NLP, and Network analysis techniques and integrating multiple heterogeneous data sources. To be specific, I look for answers to the below research questions:

(I) **How can we make use of Big Data for asthma surveillance to enable health providers to respond more promptly**? In [2] [3], I introduce a novel method of using Big Data for predicting the number of asthma-related emergency department (ED) visits in a specific area. The findings show that the proposed model can predict the number of asthma ED visits based on near-real-time environmental and social media data with approximately 70% precision. The results can be helpful for public health surveillance, ED preparedness, and targeted patient interventions.
(II) **How can we use Big Data for asthma risk factor analysis to enhance chronic disease self-management and population-level interventions**? In [4], I introduce a data-driven framework to (a) derive characteristics of asthma patients from social media data rather than traditional survey-based methods, (b) make full use of various readily available heterogeneous data sources and repurpose them to identify asthma risk factors, (c) and reveal interconnections among these risk factors and understand their relative importance. Our findings show that (1) social media data is a valuable source for delineating characteristics of self-reported asthma patients; (2) the proposed framework is very useful for a comprehensive analysis of asthma risk factors. The proposed framework can provide guidance for developing asthma management plans and population-level asthma interventions, and potentially reduce the societal burden of asthma.
(III) **After identifying smoking as one of the highest population-attributable risk factors of asthma in (II), can we use Big Data to evaluate possible substitutes like e-cigarettes**? In [5], I propose a framework to explore if Electronic Nicotine Delivery Systems (ENDS) is a safer substitute of cigarettes for asthma patients. The proposed framework collects ENDS discussion forum data, adapts techniques such as semi-supervised Medical Named Entity (MNE) recognition and the causal relationship extraction of ENDS usage and MNE. A significant reduction in asthma exacerbations is observed based on the self-reports of the asthmatic ENDS user groups on social medias. The findings of this study show that ENDS could be a valid option for asthmatic patients who have difficulties in quitting smoking.
(IV) **How can we extract health-related signals from the noisy social media data?** Social media data are notoriously noisy: first, a large percentage of data from social media, such as Twitter and Facebook, focus on advertising; second, linguistic noises are difficult for machine interpretation. Extracting health-related signals from social media data is essential for producing robust answers for above-mentioned research questions. However, there is no existing framework to extract health-related signals from social media without time-consuming and labor-intensive tasks like data annotation. Hence, I propose a comprehensive NLP and ML based framework to efficiently identify noise and extract signals from social media texts [6] [7]. The proposed framework makes a significant methodological contribution by developing a feature augmentation and sample-reweighting-based Domain Adaptation method to reduce the training effort for signal extraction by reusing previously annotated data. The experiment results show that the proposed method outperforms other baseline methods by a large margin. This work is applicable in various social media domains and can significantly improve the quality of social media datasets.

My future work will still focus on Data Science and Information System Design. In the short term, I will continue working on problems that are directly downstream of my current research, including clinical decision support and post-market surveillance of drugs. Meanwhile, all the developed models, frameworks, and design principles are not limited to asthma. They can be generalized and

used for other chronic conditions such as obesity, type 2 diabetes, cardiovascular disease, and cancer. In the long run, my research aims to bridge the gap between Data Science and healthcare analytics. Applying Big Data analytics to healthcare is not only important to Design Science, but equally important to Behavior and Social sciences. A lot of studies highlight the importance of both social (e.g., social position or status; quality of social relationships) and psychological exposures (e.g., life stress; extreme emotional such as anger or fear) in the exacerbation of chronic illness. Yet, there are insufficient studies that investigate these research problems systematically and quantitatively. Healthcare-related Big Data analytical tools have the potentials to leverage data from large-scale longitudinal sources for population level chronic disease prevention, as well as to capture trends and propose models for individual-level proactive self-management. Combining Big Data analytics with Behavior and Social sciences will further help us understand the mechanisms of chronic disease development and progression.

In summary, healthcare-related Big Data analytics remains an open research area with many challenging issues. From the perspective of Design Science, my current work advances the analysis of Big Data in healthcare by designing NLP, ML, and Network based artifacts. The findings have significant implications for researchers, health providers and patients with chronic disease. Moreover, there are adequate resources and funding opportunities in this research area. First, social media, real-time sensors, and the Internet based data are readily available and inexpensive in use. Second, healthcare-related research has been a worldwide focus for governments, international organizations, and scientists in decades. Annually, more than $90 billion is spent on public and philanthropic health research globally. The largest funder is the US National Institutes of Health ($32.3 billion annually) [8]. It is of little doubt that this research area will remain active in the future.

References

[1] CDC, “National Center for Chronic Disease Prevention and Health Promotion,” 2016. [Online]. Available: https://www.cdc.gov/chronicdisease/. [Accessed: 17-Apr-2017].

[2] S. Ram, W. Zhang, M. Williams, and Y. Pengetnze, “Predicting Asthma-Related Emergency Department Visits Using Big Data,” *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 4, pp. 1216–1223, Jul. 2015.

[3] W. Zhang, S. Ram, M. Burkart, and Y. Pengetnze, “Extracting Signals from Social Media for Chronic Disease Surveillance,” in *Proceedings of the 6th International Conference on Digital Health Conference*, Montreal, Quebec, Canada, 2016, pp. 79–83.

[4] W. Zhang and S. Ram, “A Comprehensive Analysis of Risk Factors for Asthma: Based on Machine Learning and Large Heterogeneous Data Sources,” [Under Review].

[5] W. Zhang and S. Ram, “Are Electronic Cigarettes a Safer Substitute for Cigarettes for Asthma Patients?,” [Working Paper].

[6] W. Zhang and S. Ram, “A Comprehensive Methodology for Extracting Signal from Social Media Text Using Natural Language Processing and Machine Learning,” in *the Proceedings of the 25th Workshop on Information Technologies and Systems (WITS)*, Dallas, Texas, USA, 2015.

[7] W. Zhang and S. Ram, “Domain Adaptation for Signal Extraction from Large Social Media Datasets,” the Conference on Information Systems and Technology (CIST), 2017, [Under Review].

[8] R. F. Viergever and T. C. C. Hendriks, “The 10 largest public and philanthropic funders of health research in the world: what they fund and how they distribute their funds,” *Health Research Policy and Systems*, vol. 14, p. 12, Feb. 2016.