Growing Diversity in Data Science: Shared Lessons from Clinical Trials

Robert D. Simari, MD
Department of Cardiovascular Medicine
University of Kansas School of Medicine
University of Kansas Medical Center

The demographic nature of western society is rapidly changing. In the United States the population is aging and becomingly increasingly diverse (1). Next year, there will not be a majority race among those under 18. By 2060 there will be no majority within the entire US population. The implications of these changes are enormous and the academic enterprise will not be spared. The work of academia and the work force of academia will be forever changed within these ongoing social changes. Data science has the potential to alter the fundamental framework of biomedicine. Machine learning and artificial intelligence have the capacity to identify mechanisms and associations that may lead to innovations in disease prevention and therapy. Yet data science must evolve with the social changes underway.

In the field of clinical trials the diversity of the investigators and the subjects have major impact on the conduct and applicability of the trials. Unlike the experimental nature of clinical trials, data science is observational in nature. Yet decisions in how data sets are generated and analyzed are made by individuals and groups of individuals. The thoughts and actions of each of those individuals is based on their life experiences. As such, diversity of the investigative team can impact the conduct and outcome of data science. In this paper, I will extend the importance and challenges of diversity in clinical trials and apply it to the new field of data science.

In 2018, an editorial in Nature was published entitled "When will clinical trials reflect diversity"(2). In this essay the authors demonstrate the lag between social change and inclusiveness in subjects in clinical trials. In spite of federal requirements and expectations, clinical trial subjects are mainly white and male. Yet, it is widely understood that social determinants of health are the major drivers of health in the United States. As such, exclusion of diverse populations in clinical trials may result in underpowered studies whose results may not be generalizable to the broader population. Further, the exclusion of large segments of our population from these therapeutic trials is not just.

The reasons for a lack of diversity in clinical trials find their roots in social injustices throughout our country's history. Historic injustices in medical experimentation have led to significant lack of trust among diverse populations. This lack of trust has led to challenges in the recruitment of underrepresented communities. Furthermore, there are social and financial barriers that are inherent to inclusion in trials. Missing work, lack of transportation and limited exposure to available trials are often cited as barriers. Progress to overcome this lack of diversity has been slow in spite of the multiple drivers. One particular area of focus is to address the primacy of diversity of the investi-

gative team. Diverse investigators will be more suited to design trials and approaches that favor more inclusion and strategies to recruit broader populations. The need for diversity of the investigative team may also be paramount in developing diversity in data science.

Why is it important to have diverse investigative teams in data science? First, the generation of data bases may be biased by the teams that generate them. Having a diversity of thought, opinion and background may limit known and unconscious bias that can affect the data sets. Second, the analysis of data can be similarly biased without a diversity of thought. Third, the critical social and medical issues that impact diverse populations who bear the greatest burden of social determinants can be impacted through data science. The likelihood that data science focuses on these issues will be enhanced if diverse investigative teams are developed. Finally, the population of majority students in the pipeline will certainly decrease. If diverse groups are not considered for advanced training, the overall workforce in data science will suffer.

The rate of societal changes in diversity is noticeably greater than changes in diversity in academia. Data from the National Science Foundation demonstrate that the gap for underrepresented populations is greater for higher academic degrees than entry level degrees, is greater for men than women and is greater for black and African-Americans than for Hispanics and Latinos (3). Furthermore, these gaps are greater in the sciences associated with data science, mathematics and engineering. These gaps and the rates at which they are lessening do not suggest that the current problem in diversity will be self-limited.

So what strategies might be successful in developing diverse investigative teams in data science? Unfortunately, the challenges faced by the field of clinical trials are applicable to data science as well. Prepared doctoral graduates in advanced fields of science and technology are required. For clinical trials the fields are medicine, nursing and biomedical studies. In data science it is mathematics, engineering, medicine, computer science and statistics. As discussed above, major challenges remain with a lack of diversity in these disciplines.

As the trends in diversity demonstrate, passive approaches will not be successful towards the goal of diverse investigative teams. Intentional and continuous efforts will be required. Additionally these efforts must be focused at the time in which students begin to develop aptitudes for STEM fields. In my opinion the cornerstones of such programs include the following:

1. Coordination among centers of higher education and communities of underrepresented minorities to provide valued and culturally competent programs to support attainment of educational milestones.
2. Exposure of diverse students to careers in data science throughout their pregraduate careers preferably with diverse role models, mentors and sponsors.
3. Coordination with governmental and non-governmental organizations to address social determinants of health, many of which act as social determinants of education and achievement.

At the University of Kansas Medical Center, we have engaged our Kansas City, Kansas community through a series of programs that attempt to meet the cornerstones defined above. These programs start at the earliest stage of development

with Project Eagle, a large university run Head Start program supporting some of the neediest and youngest members of our community. Our faculty are engaged in multiple programs throughout the public K-12 school including a successful grant which funds STEM curriculum development for our local high schools and a summer and Saturday science academy. Finally, through the advocacy of our faculty and staff, we have enabled building of sidewalks and grocery stores in food deserts in our community.

For students interested in health based careers, we have a formal shadowing program, a post bac degree for underrepresented minorities that includes a guarantee of medical school admission and transition to a prematriculation program for successful graduates. Taken together, this series of programs creates an important mechanism for students to meet the educational demands of STEM careers.

The field of data science has the potential to revolutionize academia and our society. Yet it is subject to the unconscious biases currently present in both. We must strive to broaden the diversity of investigators who participate in data science in order to avoid the pitfalls currently present in the field of biomedical clinical trials. This will require intentional and consistent efforts throughout the educational hierarchy and interspersed throughout our communities. The future of data science is dependent on those efforts.

References

1. U.S. Census Bureau (2017). 2017 National Projections Tables. https://www.census.gov/data/tables/2017/demo/popproj/2017-summary-tables.html Accessed 9/5/19.
2. Knepper TC, McLeod HL. When will clinical trials reflect diversity. Nature. 2018 May; 557(7704):157-159. doi: 10.1038/d41586-018-05049-5.
3. National Science Foundation. Women, minorities, and persons with disabilities in science and engineering. Arlington, VA: NSF 17-310, Jan 2017. https://www.nsf.gov/statistics/2017/nsf17310/digest/about-this-report/. Accessed 9/5/19.