Machine Learning in Population and Public Health

Vishwali Mhasawade
vishwalim@nyu.edu
New York University

Yuan Zhao
yuan.zhao@nyu.edu
New York University

Rumi Chunara
rumi.chunara@nyu.edu
New York University

ABSTRACT

Research in population and public health focuses on the mechanisms between different cultural, social, and environmental factors and their effect on the health, of not just individuals, but communities as a whole. We present here a very brief introduction into research in these fields, as well as connections to existing machine learning work to help activate the machine learning community on such topics and highlight specific opportunities where machine learning, public and population health may synergize to better achieve health equity.

ACM Reference Format:
Vishwali Mhasawade, Yuan Zhao, and Rumi Chunara. 2020. Machine Learning in Population and Public Health. In *ACM Conference on Health, Inference, and Learning (ACM CHIL '20).*

1 POPULATION AND PUBLIC HEALTH

Population and public health are approaches to health research and practice which aim to to understand what makes and keeps people healthy [52]. The major underpinning principle for this approach is that of *health equity*, defined as *"Minimizing avoidable disparities in health and its determinants – including but not limited to health care – between groups of people who have different levels of underlying social advantage or privilege, i.e., different levels of power, wealth, or prestige due to their positions in society relative to other groups"* [4]. Thus this guiding principle necessitates a focus on determinants, antecedents and other factors related to health outside the hospital. As further described by the socioecological model of health (Figure 1), a cornerstone of these domains which provides a conceptual model to illustrate how the health of an individual is affected by multiple factors operating at different levels in a hierarchy, these multi-level factors include public policies at the national and international level, availability of health resources within a neighborhood, community behavior, and ultimately the habits and behavior of individuals [6]. Understanding the complex interactions between individuals and their environments is crucial to realize not just how the health of the high-risk individuals, for example, those suffering from cardiovascular disease risk, can be improved but also what policies would benefit the community as a whole, such as will introducing healthier food options in a neighborhood help people to improve their diet [55]?

ACM CHIL '20, July 23–24, 2020, Toronto, ON, Canada

Figure 1: The socioecological model of health.

1.1 Potential impact

Realizing these complex interactions and identifying their effect on health outcomes remains one of the most prominent challenges as impact of doing so would be very large. For example, the inequalities in life expectancy for women with different income levels in the United States are large and growing [42]. As well in the United States, social factors account for 25-60% of deaths in any given year according to results from various meta-analyses [24], and these factors play a significant role in prevention of the five leading causes of death (diseases of the heart, cancer, unintentional injuries, cerebrovascular diseases, and chronic lower respiratory diseases) [9]. Worldwide, 80% of the growing burden of noncommunicable diseases could be prevented through modifying behaviors such as reducing tobacco/alcohol or fat and salt consumption, promoting physical activity and improving environmental conditions such as air quality and urban planning [59].

Although in the past decade statistical and machine learning approaches have made considerable progress in automating clinical tasks, here we aim to illustrates where and how such techniques can be useful in a more holistic view of health. Towards this, we summarize existing challenges in public and population health and related opportunities. The outline of the document is as follows: we introduce the data used in population and public health studies in Section 2, followed by approaches for analyzing risk factors and disparities in Section 3, and finally, present some directions for future work that leverages synergies in machine learning and population and public health in Section 4.

2 DATA IN PUBLIC HEALTH

Considering that health is affected by individual-level and community-level attributes, a large focus of work is on identifying and understanding all of these potential factors and how they affect health outcomes. Naturally, this involves assessing not just biological factors but also social factors such as income, education, and socioeconomic status (SES). The World Health Organization (WHO)

defines these social determinants of health (SDoH) as *"conditions in which people are born, grow, live, work and age", such circumstances are shaped by the distribution of power, money, and resources, not just at the local level but also global and national levels* [58]. It is important to note that these factors can come into play and have effect throughout the life course or across generations [8]. For example, adverse childhood experience of parents was found to be associated with higher odds of poor health of the child [30]. Thus it is paramount to look beyond the individual's immediate risk factors and assess such inter-generational effects as well as the social, economic, and cultural environments that an individual identifies with. We next present some of the approaches to identifying and assessing social factors. Following we describe, through the deep literature on social determinants, common approaches for their measurement, and we distill challenges which can illuminate potential opportunities for machine learning.

2.1 Measuring social determinants of health

Traditional approaches to data collection in public health involve aggregation across multiple levels. Individuals often report data to public health practitioners and healthcare workers, which is then aggregated by local officials and forwarded to health ministries at the national level and finally analyzed at the international level by organizations like World Health Organization, United Nations. This results in robust, denominator-based data available from public health or governmental organizations, such as the national health and nutrition examination survey (NHANES) [18] or the behavioral risk factor surveillance system (BRFSS) [54]. These types of data systems also aim to capture different indicators which compose each construct (e.g. housing quality can be measured through data on rental status, sanitation status, crowding, indoor air quality, etc.) [28]. However, there is also a loss of critical information at an individual level due to aggregation as well as temporal delays in the process. Accordingly, the rising ubiquity in technologies such as smartphones and social media sites like Twitter or Instagram has come to the forefront, making it possible to access high-resolution data that does not suffer from recall or information biases, and to capture information across daily behavioral patterns [11, 36, 62].

These new data sources provide opportunity for measuring social determinants, but also come with their own challenges. After over a decade of their use in health studies, it is now possible to distill common challenges. Beyond data privacy, which we highlight but do not discuss here at length as there are several other reports which focus on it [40, 46], one of the most prominent issues is understanding the study's denominator, i.e., who all were included in study [11]. Moreover, since studies are often focused on specific groups in the community, this raises internal and external validity issues. First, it is important to understand what the measured variables are indicative of, and what constructs they represent. For example, recent work on how individual-level syndromic reports or passive data from individuals relates to microbiologic confirmation [14], aims to address early challenges highlighted in using data such as Google searches to predict influenza [29]. Research on the data generating process of new digital data sources is also imperative to understand what the data represents, and why data is being shared by individuals [37]. Second, can the results be extended to other situations, groups, or events; which sub-populations are representative of the hypothesis [38]? To understand this phenomenon, a recent study analyzed the characteristics of different surveillance approaches (such as the questions asked, time of data collection) for influenza and their effect on predictive performance. Even if similar syndromic data is collected across all the approaches, there are considerable differences in the predictive value of the syndromic data [12], highlighting that the same type of data can represent different factors across studies (in this case the mode of data collection can affect the specificity/sensitivity of respiratory infection syndromic data). Although there are significant differences in sample representation and predictivity across studies, it is crucial to understand external validity as it provides a way to understand population-level characteristics that remain invariant across studies [11, 37].

Challenges in measuring social determinants echo recent work in the computer science literature. Holstein et al. interviewed data scientists and synthesized findings around the importance of better data for improving the performance of machine learning models opposed to model development [25]. Further, mapping from the construct to the observed space can itself be a place where bias can enter [19], and thus special attention to which data and from whom it is gathered, should be highlighted in health research.

2.2 Integrating social determinants of health in machine learning models

Beyond work capturing social determinants using machine learning from person-generated sources [1, 48], a recent systematic review analyzed how social determinants have been used to study the risk factors of cardiovascular diseases [63]. While common social determinants like age, gender, race, income, and education have been analyzed for estimating cardiovascular risk, most commonly the factors considered are at the individual-level even though research has clearly shown that community-level factors such as a person's neighborhood's overall income can also affect their disease risk. [63]. Moreover, most studies to-date using machine learning models have involved associative studies. However, it is essential to understand the complex causal pathways between the social factors and the health outcomes in order to design effective interventions. Thus, there is a need to look beyond familiar data sources such as electronic health records (EHR), and develop approaches for evaluating the causal underpinnings in the data that captures these multi-level factors. Recognizing and understanding these social determinants of health across high dimensional multi-modal sources is an area which may benefit through the use of machine learning.

2.3 Are social variables intervenable?

Sometimes the use of social variables in causal models is restricted, under the premise that they are non-manipulable, or not intervenable [21]. Moreover, causal methods often assume stable unit treatment value (SUTVA). This implies that there is no interference and only one version of treatment, but this is often impossible with the complex interaction between social determinants. One proposed solution for this is to manipulate downstream mediators, such as encouraging children to read in order to increase their cognitive

ability. But in order to make structural changes, we still need to address the root cause of disparities, and also aim at intervening on upstream social determinants, such as improving SES [31].

At the same time, quantitative health researchers have aimed to understand how structural factors relate to health while also keeping in mind what is possible to intervene. One example of this is a simulation study aimed at understanding factors related to reducing the prevalence of chronic illnesses. While it was postulated that multiple social factors like social cohesion, jobs/income, education level, individual behavior, housing and healthcare interventions can lead to reduction in chronic illness, the simulation study narrowed down which factors were related based on data from their setting. They were also able to quantify the intervention needed for each social factor in order to lower chronic illness prevalence [34].

3 TYPES OF HEALTH TASKS

Guided by the principle of health equity defined above, we see that a broad focus on multi-level aspects related to health in order to understand as well as intervene on is essential. Accordingly, we developed a taxonomy to consolidate types of tasks. We provide examples of where machine learning has been used in each type of task, and where further innovation in each of these is possible as well. The taxonomy includes: 1) identification of factors (biological, environmental, social, etc.) and their relation to health, 2) design/evaluation of interventions and policies on health, 3) prediction of outcomes, and 4) allocation of resources at the individual or group level. We briefly describe examples of such tasks below, full discussion on the health taxonomy can be found at https://chunaralab.github.io/MLPH.

(1) *Identification of factors.* Learning what contributes to health outcomes is a common theme across biological, social and other factors, and a good opportunity for learning from data which is a fundamental aspect of machine learning. The concept of identification comes into play across a broad range of studies in health, from learning biological mechanisms [7], assessment of treatment effects [33] to epidemiological studies of spatial factors [2].
(2) *Design of interventions.* Interventions can occur at multiple levels and through different mediums, providing diversity in this category as well. For example, group based intervention programs are one of the means for reducing substance abuse by reinforcing positive behavior. With the advent of digital data available, social networks can be used for designing interventions and targeting high risk individuals in proximity with already exposed users. For example, Rahmattalabi et al. [49] present an influence-based partitioning of social networks to identify high impact intervention groups.
(3) *Prediction of outcomes.* While predicting treatment effectiveness [27], mortality risk [51], hospital readmission [20], and disease prognosis [15] are some tasks well studied in machine learning, predicting risk scores with clinical algorithms while mitigating health disparities as well as prediction of outside-hospital events are still crucial challenges [57].
(4) *Allocation of resources.* Resource allocation is a well studied problem in artificial intelligence. For example, in order to ensure equity in access to resources Snyder et al. [53] suggest an allocation system combining a medical priority score and a geographic feasibility score to facilitate organ allocation across geographical boundaries. However, there is further opportunity to consider the types of resources and attributes which are considered in allocation, with a health equity lens.

3.1 Causal approaches to understand mechanisms between social determinants and health

Causal approaches are critical for understanding the mechanism between different social factors and health outcomes. Causal methods provide a way to incorporate prior knowledge about mechanisms into modeling, and also to identify what is intervenable. For example, it is known that the social construct of 'race' has direct effects on the health of individuals belonging to the disadvantaged group. To improve the health of the marginalized individuals, it may be argued that *race* cannot be intervened upon and there is nothing that can be done to *race* to improve the existing scenario. However, the social construct can be decomposed into several interacting factors like parents' genetics, family culture, social perspectives about appearance, the early life socio-economic conditions, and late-life socio-economic conditions. While some of the factors like parents' genetics cannot be intervened upon, it is still possible to mitigate health disparities by focusing on intervenable factors like socio-economic conditions [56].

Another aspect of modeling social determinants apart from decomposing social constructs into interacting factors is realizing the different pathways and interaction effects (i.e. learning the data generating process). For example, an analysis of the health behavior of individuals on social media website Instagram revealed that the immediate environment comprising of the food resources in the neighborhood as well as the social network of an individual (the profiles being followed and interacted with) affect what individuals post on Instagram about dining [36]. What is interesting here is the focus on learning the mechanism through which the social environment has an effect on the health behavior, more broadly than existing public health knowledge. While existing work in public health has shown there is a direct effect of the availability of resources in the proximity of an individual and their health [41], this work augments current understanding to include factors from the online environment and understanding the mechanism between the online and social environment and in turn its effect on what an individual posts online.

Causal methods are critical in models, including when considering social variables. A natural experiment across various states in the United States (with varying compulsory education laws) analyzed the impact of education level on cardiovascular disease risk. A postulated causal graph with multiple pathways from education level to cardiovascular disease risk is represented in Figure 2. A simple association method concludes that education level decreases cardiovascular disease risk across all risk factors such as BMI, cholesterol, smoking, and depression. On the other hand, a causal approach, known as instrumental variables (IV), presents that education does not improve BMI and cholesterol risks [23]. This

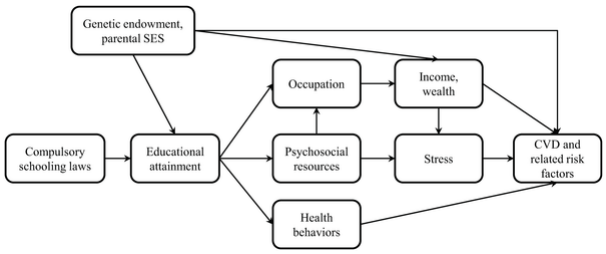

Figure 2: Conceptual model linking educational attainment level with cardiovascular disease (CVD) [23].

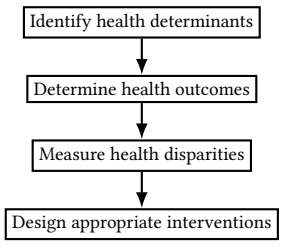

Figure 3: Pipeline for detecting health disparities

can be attributed to the sedentary lifestyle resulting from improved income sources correlated with increased education levels. It is thus essential to identify the various interacting factors comprising a social construct and estimate the effects of social determinants via appropriate causal methods.

3.2 Pitfalls with 'proxies' in health modeling

Although it is necessary to include and analyze the several comprising factors of social constructs like 'race,' it is often unknown what the comprising factors are, how they interact, and how to model them. An inaccurate understanding of variations within and between groups, along with difficulty in attaining the relevant social factors have lead to the use of *known* social constructs like race as *proxies* for these unknown, unmeasured factors. A recent study highlighted several clinical tools across cardiology, nephrology, obstetrics, and many other specialties, which all use race while estimating risk factors, and how this use severely compromises the health of marginalized individuals [57]. A specific example of this is in the common way that kidney function is estimated using glomerular filtration rate (GFR) (estimated glomerular filtration rate) which uses serum creatinine levels and the race of the individual. Serum creatinine is the waste product in the blood resulting from muscle activity, is absorbed from the blood by the kidneys. However, under abnormal kidney function, the level of serum creatinine increases in the blood. Another reason for the increased serum creatinine level is higher muscle mass, which is attributed as the reasoning for considering the race factor. However, further research and understanding of the construction of this equation have illuminated that the use of race is not appropriate. It is not an accurate measure of the proxy, and furthermore results in delayed and disparate treatments to patients identifying as *Black* individuals [16, 22, 32]. Discussion of these issues has led to the elimination of race considerations in calculating eGFR in several hospitals around the United States [44, 64]. How any relevant genomic variation can be assessed and reported without stratifying populations based on factors like race and ethnicity is a challenge to be addressed [3].

3.3 Multiple axes of disparities and intersectionality

A health disparity/inequality is a particular difference in type of health where disadvantaged social groups like women, poor, racial/ethnic minorities experience greater health risks [4]. Such disparities are reflective of social oppression and its influence on the health of disadvantaged individuals. It is essential to identify what leads to disparate health outcomes in order to design interventions to mitigate disparities and improve the health of high risk populations. This involves multiple tasks involving what factors affect health including both individual-level and social factors as mentioned before [35, 36], measuring the health outcomes [10, 45], measuring disparities across social groups [5, 47, 50], and finally designing policies to mitigate the disparities [50]. The steps towards mitigating disparities are summarized in Figure 3.

We now present a specific challenge in measuring disparities across groups of people. Health disparities are often measured by considering the average health outcomes of individuals identifying to a specific social construct say *race*. For example, is the psychiatric readmission rate the same across racial groups [10]. However, measuring health disparities as averages is not enough to identify the narrower sub-populations still suffering from the burden. A longitudinal study of prenatal care access for childbearing women in California across 1994-1995 and 1999-2001 reveals that the federal and state policies within these periods, on average, improved the health of pregnant women. However, disparities continued to exist across income groups even with the introduction of policies [4]. Thus, it is vital to measure differences across multiple social axes such as race, income, and education and primarily focus on individuals existing at the intersections of social disadvantages [13].

One approach for addressing the disparities at intersections of social factors is a multilevel approach known as MAHIDA (Multilevel Analysis of Individual Heterogeneity and Discrimination Accuracy) [17]. It involves decomposing the total variance into a) between-strata variance focusing on identifying and assessing disadvantaged social groups, and b) within-strata variance which aims to identify individuals within a social group that are at added disadvantaged as compared to other members of the group. The approach presents several advantages over fixed-effect models that include interaction terms for multiple sensitive attributes. These include restricting the parameter growth to linear compared to geometric and adjusting for the sample sizes of the intersectionalities. The approach is consistent with the field of eco-epidemiology that cautions against aggregation. Studies have also focused on incorporating information across multiple environments to learn population-level characteristics, especially when subgroups are underrepresented across individual studies [37].

3.4 Health disparities and algorithmic fairness

The growing use of machine learning methods in healthcare has raised the question of whether the model outcomes are discriminatory based on variables like race and ethnicity that are often used in constructing predictions [10, 37]. A recent literature in machine learning and statistics aims on ensuring that model outcomes are not discriminatory towards individuals who have the same *merit* [26, 39]. For example, multiple fairness definitions like equality of opportunity, demographic parity have been suggested as approaches to restrict model outcomes to ensure fairness. However, given the complex causal relationships between the biological, social and environmental factors that lead to disparities in health outcomes [35] outlined here, questions remain regarding such advances in algorithmic fairness and how they may interface with health disparities [19, 26, 35]. Moreover, assessing disparate health outcomes becomes challenging when the underlying causal mechanisms are not known. We briefly outline specific challenges regarding algorithmic fairness and health disparities.

Algorithmic design goals. First, it is crucial to ensure that the algorithm design supports the goal of *health equity*. For example, Obermeyer and Mullainathan [45] present that racial disparities in risk scores are a result of considering financial cost expenditure as a proxy for health care needs. It is therefore essential to be aware of such proxies and also assess disparities through different causal pathways, namely 1) direct, where the social construct has a direct impact on the outcome, and 2) through indirect causal pathways which provide an opportunity for performing interventions [60].

Unexplained variance resulting from proxies. Consider the scenario represented in Figure 4 where we are concerned with predicting the health outcome, say risk for cardiovascular disease using the protected attribute, and clinical conditions. We also want to restrict models outcomes to be fair using some fairness metric like demographic parity. Even if we are successful in ensuring that the risk is fair with respect to the *racial identity*, we are still left with unexplained variance for the social components of race. As we illustrate above, race is a social construct, and it would be unclear if such a model accounts for variances in the true mechanisms which race is acting as a proxy for. For example, do equal risk scores for Black and non-Black patients also ensure that there is no disparity across, say lower-income Black patients vs. higher-income Black patients. Thus, there is a need to both identify intersectional social groups as well as underlying causal mechanisms, and ensure algorithmic fairness for the same.

Equity vs. in-sample fairness. Since data input for the algorithms may not be representative of the population for which decision might be made, it is crucial to be aware of the sub-populations included in the study [43, 61], and mitigate disparate outcomes for under-represented sub-populations [37]. Figure 5 represents a common case in healthcare where different individual factors are used to make algorithmic decisions for patients. A fair algorithmic restricting the model decisions in favor of P still suffers from population unfairness even if it achieves in-sample fairness. This is worsened by considering the insurance type of the individual, which leads to an association between the said insurance and health outcomes. It is important to note that often individuals are unable to access care, and remain at higher risk. Thus, it is imperative to consider if the aims of our efforts will alleviate and not exacerbate health inequities.

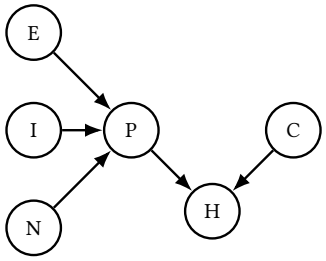

Figure 4: Perceived protected attribute P is composed of several social factors such as education E, income level I, and neighborhood characteristics N. Health outcome H is to be predicted using the clinical variables C, and protected attribute P while ensuring that the model prediction is fair with respect to P.

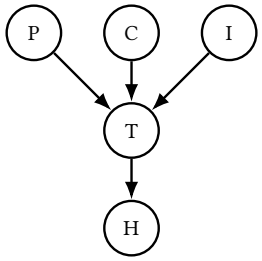

Figure 5: Perceived protected attribute P, clinical variables C, and the insurance type I used to determine the treatment T, and ultimately assess the health outcomes H.

4 CONCLUSION

Health equity is a vast concept, and one that requires understanding and assessment in an ongoing manner, as factors related to health are shaped and change. Through the summary of existing work in population and public health, specific discussion of the importance of social determinants and challenges in their measurement and incorporation into causal models, we have synthesized the many open areas for *machine learning*, to advance and build on research and practice in this area. The taxonomy helps lay out different areas in health where machine learning has and can play an important role; identifying factors related to health outcomes, design/evaluation of interventions and policies, prediction of outcomes and allocation of resources. Finally, we also discuss how the important growing research on algorithmic fairness interfaces with health disparities. Full discussion on the health taxonomy and related topics can be found at https://chunaralab.github.io/MLPH. In sum, there are many opportunities to build on the deep body of work on health equity. We hope that this work serves to inform and activate the machine learning community on these critical topics.

5 ACKNOWLEDGMENTS

We thank Dr. Stephanie Cook and Harvineet Singh for helpful discussions on this topic. The work was partially supported under National Science Foundation grant 1845487.

REFERENCES

[1] M. Akbari and R. Chunara. Using contextual information to improve blood glucose prediction. *Machine Learning for Healthcare*, arXiv preprint arXiv:1909.01735, 2019.
[2] S. Bhatt, P. W. Gething, O. J. Brady, J. P. Messina, A. W. Farlow, C. L. Moyes, J. M. Drake, J. S. Brownstein, A. G. Hoen, O. Sankoh, et al. The global distribution and burden of dengue. *Nature*, 496(7446):504–507, 2013.
[3] V. L. Bonham, E. D. Green, and E. J. Pérez-Stable. Examining how race, ethnicity, and ancestry data are used in biomedical research. *Jama*, 320(15):1533–1534, 2018.
[4] P. Braveman. Health disparities and health equity: concepts and measurement. *Annu. Rev. Public Health*, 27:167–194, 2006.
[5] P. A. Braveman, S. A. Egerter, C. Cubbin, and K. S. Marchi. An approach to studying social disparities in health and health care. *American Journal of Public Health*, 94(12):2139–2148, 2004.
[6] U. Bronfenbrenner. Toward an experimental ecology of human development. *American psychologist*, 32(7):513, 1977.
[7] S. Burgess, C. N. Foley, and V. Zuber. Inferring causal relationships between risk factors and outcomes from genome-wide association study data. *Annual review of genomics and human genetics*, 19:303–327, 2018.
[8] N. Cable. Life course approach in social epidemiology: an overview, application and future implications. *Journal of epidemiology*, page JE20140045, 2014.
[9] CDC. Up to 40 percent of annual deaths from each of five leading us causes are preventable. *Atlanta, GA: Centers for Disease Control and Prevention*, 2014.
[10] I. Y. Chen, M. Agrawal, S. Horng, and D. Sontag. Robustly extracting medical knowledge from ehrs: A case study of learning a health knowledge graph. In *Pac Symp Biocomput*, pages 19–30. World Scientific, 2020.
[11] R. Chunara, L. E. Wisk, and E. R. Weitzman. Denominator issues for personally generated data in population health monitoring. *American journal of preventive medicine*, 52(4):549–553, 2017.
[12] R. Chunara, A. Plymoth, and L. Martin. Diversity in surveillance data: implications for infectious disease forecasting models. *Under review*, 2020.
[13] K. Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.*, page 139, 1989.
[14] A. R. Daughton, R. Chunara, and M. J. Paul. Comparison of social media, syndromic surveillance, and microbiologic acute respiratory infection data: Observational study. *JMIR Public Health and Surveillance*, 6(2):e14986, 2020.
[15] T. M. Dugan, S. Mukhopadhyay, A. Carroll, and S. Downs. Machine learning techniques for prediction of early childhood obesity. *Applied clinical informatics*, 6(03):506–520, 2015.
[16] N. D. Eneanya, W. Yang, and P. P. Reese. Reconsidering the consequences of using race to estimate kidney function. *Jama*, 322(2):113–114, 2019.
[17] C. R. Evans, D. R. Williams, J.-P. Onnela, and S. Subramanian. A multilevel approach to modeling health inequalities at the intersection of multiple social identities. *Social Science & Medicine*, 203:64–73, 2018.
[18] N. C. for Health Statistics (US). *Plan and operation of the third National Health and Nutrition Examination Survey, 1988-94.* Number 32. National Ctr for Health Statistics, 1994.
[19] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.
[20] P. Galiatsatos, A. Follin, N. Uradu, F. Alghanim, Y. Daniel, S. Saria, J. Townsend, C. Sylvester, A. Chanmugam, and E. Chen. The association between neighborhood socioeconomic disadvantage and readmissions for patients hospitalized with sepsis. In *C94. The Impact of Social Determinants in Pulmonary and Critical Care*, pages A5569–A5569. American Thoracic Society, 2019.
[21] C. Glymour and M. R. Glymour. Commentary: race and sex are causes. *Epidemiology*, 25(4):488–490, 2014.
[22] V. Grubbs. Precision in gfr reporting: Let's stop playing the race card, 2020.
[23] R. Hamad, T. T. Nguyen, J. Bhattacharya, M. M. Glymour, and D. H. Rehkopf. Educational attainment and cardiovascular disease in the united states: A quasi-experimental instrumental variables analysis. *PLoS medicine*, 16(6):e1002834, 2019.
[24] H. J. Heiman and S. Artiga. Beyond health care: the role of social determinants in promoting health and health equity. *Health*, 20(10):1–10, 2015.
[25] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2019.
[26] M. Kasy and R. Abebe. Fairness, equality, and power in algorithmic decision making. Technical report, Working paper, 2020.
[27] N. Kreif, R. Grieve, I. Díaz, and D. Harrison. Evaluation of the effect of a continuous treatment: a machine learning approach with an application to treatment for traumatic brain injury. *Health economics*, 24(9):1213–1228, 2015.
[28] S. V. Kusnoor, T. Y. Koonce, S. T. Hurley, K. M. McClellan, M. N. Blasingame, E. T. Frakes, L.-C. Huang, M. I. Epelbaum, and N. B. Giuse. Collection of social determinants of health in the community clinic setting: A cross-sectional study. *BMC Public Health*, 18(1):550, 2018.
[29] D. Lazer, R. Kennedy, G. King, and A. Vespignani. The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.
[30] F. Lê-Scherban, X. Wang, K. H. Boyle-Steed, and L. M. Pachter. Intergenerational associations of parent adverse childhood experiences and child health outcomes. *Pediatrics*, 141(6):e20174274, 2018.
[31] C. Lehman. Addressing social determinants of health. *Physical Therapy in Motion*, 2019.
[32] A. S. Levey, S. M. Titan, N. R. Powe, J. Coresh, and L. A. Inker. Kidney disease, race, and gfr estimation. *Clinical Journal of the American Society of Nephrology*, 2020.
[33] S. Lodi, A. Phillips, J. Lundgren, R. Logan, S. Sharma, S. R. Cole, A. Babiker, M. Law, H. Chu, D. Byrne, et al. Effect estimates in randomized trials and observational studies: comparing apples with apples. *American journal of epidemiology*, 188(8): 1569–1577, 2019.
[34] A. Mahamoud, B. Roche, and J. Homer. Modelling the social determinants of health and simulating short-term and long-term intervention impacts for the city of toronto, canada. *Social science & medicine*, 93:247–255, 2013.
[35] M. D. McCradden, S. Joshi, M. Mazwi, and J. A. Anderson. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health*, 2(5):e221–e223, 2020.
[36] V. Mhasawade, A. Elghafari, D. T. Duncan, and R. Chunara. Role of the built and online social environments on expression of dining on instagram. *International journal of environmental research and public health*, 17(3):735, 2020.
[37] V. Mhasawade, N. A. Rehman, and R. Chunara. Population-aware hierarchical bayesian domain adaptation via multi-component invariant learning. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 182–192, 2020.
[38] M. Mitchell and J. Jolley. Research design explained 5 th ed. *Victoria: Wadsworth Publisher. Moebert, J. & Tydecks, P.(2007). Power and Ownership Structures among German Companies. A Network Analysis of Financial Linkages*, 2004.
[39] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.
[40] S. J. Mooney and V. Pejaver. Big data in public health: terminology, machine learning, and privacy. *Annual review of public health*, 39:95–112, 2018.
[41] K. B. Morland and K. R. Evenson. Obesity prevalence and the local food environment. *Health & place*, 15(2):491–495, 2009.
[42] National Academies. *The growing gap in life expectancy by income: Implications for federal programs and policy responses.* National Academies Press, 2015.
[43] S. Nishtala, H. Kamarthi, D. Thakkar, D. Narayanan, A. Grama, R. Padmanabhan, N. Madhiwalla, S. Chaudhary, B. Ravindra, and M. Tambe. Missed calls, automated calls and health support: Using ai to improve maternal health outcomes by increasing program engagement. *arXiv preprint arXiv:2006.07590*, 2020.
[44] L. Nolen. *Elimination of race coefficient from eGFR calculation*, 2020. URL https://twitter.com/LashNolen/status/1276181898394558467.
[45] Z. Obermeyer and S. Mullainathan. Dissecting racial bias in an algorithm that guides health decisions for 70 million people. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 89–89, 2019.
[46] J. O'Connor and G. Matthews. Informational privacy, public health, and state laws. *American journal of public health*, 101(10):1845–1850, 2011.
[47] A. Penman-Aguilar, M. Talih, D. Huang, R. Moonesinghe, K. Bouye, and G. Beckles. Measurement of health disparities, health inequities, and social determinants of health to support the advancement of health equity. *Journal of public health management and practice: JPHMP*, 22(Suppl 1):S33, 2016.
[48] T. Quisel, D. C. Kale, and L. Foschini. Intra-day activity better predicts chronic conditions. *arXiv preprint arXiv:1612.01200*, 2016.
[49] A. Rahmattalabi, A. B. Adhikari, P. Vayanos, M. Tambe, E. Rice, and R. Baker. Influence maximization for social network based substance abuse prevention. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
[50] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.
[51] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):18, 2018.
[52] G. Rose. Sick individuals and sick populations, int. journ. *Of Epidemiology*, 14(1), 1985.
[53] J. J. Snyder, N. Salkowski, A. Wey, J. Pyke, A. K. Israni, and B. L. Kasiske. Organ distribution without geographic boundaries: a possible framework for organ

allocation. *American Journal of Transplantation*, 18(11):2635–2640, 2018.

[54] A. D. Stein, R. I. Lederman, and S. Shea. The behavioral risk factor surveillance system questionnaire: its reliability in a statewide sample. *American Journal of Public Health*, 83(12):1768–1772, 1993.

[55] D. Stern, J. M. Poti, S. W. Ng, W. R. Robinson, P. Gordon-Larsen, and B. M. Popkin. Where people shop is not associated with the nutrient quality of packaged foods for any racial-ethnic group in the united states. *The American journal of clinical nutrition*, 103(4):1125–1134, 2016.

[56] T. J. VanderWeele and W. R. Robinson. On causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology (Cambridge, Mass.)*, 25(4):473, 2014.

[57] D. A. Vyas, L. G. Eisenstein, and D. S. Jones. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms, 2020.

[58] WHO. *Closing the gap in a generation: Health equity through action on the social determinants of health: Commission on Social Determinants of Health final report*. World Health Organization, 2008.

[59] WHO. 2008-2013 action plan for the global strategy for the prevention and control of noncommunicable diseases: prevent and control cardiovascular diseases, cancers, chronic respiratory diseases and diabetes. 2009.

[60] Y. Wu, L. Zhang, X. Wu, and H. Tong. Pc-fairness: A unified framework for measuring causality-based fairness. In *Advances in Neural Information Processing Systems*, pages 3404–3414, 2019.

[61] K. Yang, J. R. Loftus, and J. Stoyanovich. Causal intersectionality for fair ranking. *arXiv preprint arXiv:2006.08688*, 2020.

[62] A. Zhan, S. Mohan, C. Tarolli, R. B. Schneider, J. L. Adams, S. Sharma, M. J. Elson, K. L. Spear, A. M. Glidden, M. A. Little, et al. Using smartphones and machine learning to quantify parkinson disease severity: the mobile parkinson disease score. *JAMA neurology*, 75(7):876–880, 2018.

[63] Y. Zhao, N. Mirin, E. Wood, V. Rajesh, S. Cook, and R. Chunara. Machine learning for integrating social determinants in cardiovascular disease prediction models: A systematic review. *In submission*, 2020.

[64] M. L. Zoler. *Dropping Race-Based eGFR Adjustment Gains Traction in US*, 2020. URL https://www.medscape.com/viewarticle/933418.