10-2021

Swimming Upstream: Getting to the Root Causes of Infant Mortality and Life Expectancy Outcomes in Cleveland, Ohio and the U.S.

Richey Piiparinen
Cleveland State University, r.piiparinen@csuohio.edu

Joshua Valdez
Rust Belt Analytica

Repository Citation

Piiparinen, Richey and Valdez, Joshua, "Swimming Upstream: Getting to the Root Causes of Infant Mortality and Life Expectancy Outcomes in Cleveland, Ohio and the U.S." (2021). *All Maxine Goodman Levin School of Urban Affairs Publications*. 0 1 2 3 1748.
https://engagedscholarship.csuohio.edu/urban_facpub/1748

SWIMMING UPSTREAM: GETTING TO THE ROOT CAUSES OF INFANT MORTALITY AND LIFE EXPECTANCY OUTCOMES IN CLEVELAND, OHIO AND THE U.S.

COMMISSIONED BY OHIO'S THIRD FRONTIER

Richey Piiparinen and Joshua Valdez [1]

CLEVELAND STATE UNIVIVERSITY The Maxine Levin College of Urban Affairs

October, 2021

[1] Richey Piiparinen is Director, Urban Theory & Analytics at Cleveland State. Joshua Valdez is Co-Founder at Rust Belt Analytica

Executive Summary

Health mostly happens outside of the hospital and the doctor's office. This reality is called "the social determinants of health" (SDOH). SDOH's impact can be understood a number of ways. This analysis shows that great local healthcare doesn't preclude poor population health, exhibited by high infant mortality and low life expectancy rates.

The analysis found that while Ohio and Cuyahoga County have world-class healthcare, they also have third-world health outcomes.

- Ohio's life expectancy (76.8) ranks 12th worst nationally, well below Midwestern peers. Cuyahoga County's life expectancy (77) ranks 41st out of 89 counties in the State.
- Black Cuyahoga County residents live shorter lives (73.6) than Whites (78.3) Hispanics (82.7) and Asians (89).
- When it comes to infant mortality rates, Ohio's infant mortality (6.97) ranks 9th worst nationally.
- Rates vary dramatically by race, with Black Ohioans (14.3) having more than double the infant mortality rates of Whites (5.1) and Hispanics (5.8).
- Cuyahoga County's infant mortality (9) is tied for second last for those Ohio counties in which the figure was calculable. The infant mortality rate for Blacks in Cuyahoga County is quadruple that of Whites and nearly triple that of Hispanics.

The analysis intended to go beyond showing *that* health disparities occur, using a novel SDOH "big" dataset to shed insight on *how* they occur. In doing so, the analysis conceptualized SDOH as either being upstream, or influenced by structural factors like class and race; midstream, or influenced by neighborhood factors like residential segregation, environmental toxins, and individual behavior; and downstream, or the prevalence of chronic disease and psychosocial stress.

Two machine learning models, or Random Forest, were run. The first model calculated the 20 highest risk factors for infant death for all counties in the Unites States. The second model calculated the 20 factors with the highest predictive power on life expectancy for all census tracts in Cuyahoga County. The results explain what factors predict high infant mortality rates between counties in the U.S. and what factors predict life expectancy rates between neighborhoods within Cuyahoga County. These factors are varied and range from the percent of knowledge workers in a neighborhood, the proximity to five star Yelp establishments, foreclosures, sheriff sales, car volume and proximity and volume, and ground-level Ozone.

In all, the analysis calls for a methodological and conceptual approach in which SDOH researchers and practitioners are "swimming upstream" to address the root causes of health disparities from a policy standpoint, while continuing to tackle midstream and downstream factors through behavioral- and neighborhood-based intervention. To do this, the authors suggest implementing more data science practices into the social science field, an interdisciplinary movement termed "computational social science"[2].

(Note: A Companion piece to this analysis can be found in the October 28th edition of the *Cleveland Plain Dealer* in the op-ed "Use Rescue Plan Act dollars to rescue Cleveland's health"[3] by Richey Piiparinen.)

[2] https://science.sciencemag.org/content/369/6507/1060
[3] https://www.cleveland.com/opinion/2021/10/use-rescue-plan-act-dollars-to-rescue-clevelands-health-richey-piiparinen.html

Introduction: SDOH and The Holon of Health

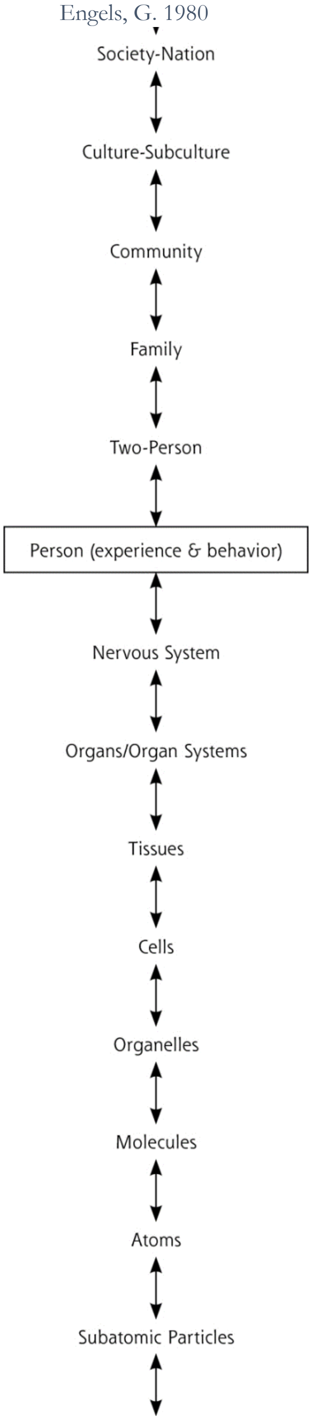
In the Summer of 1968, author-philosopher Arthur Koestler assembled a who's who of thinkers for a symposium meant to push back against "the insufficient emancipation of the life sciences from the mechanistic concepts of nineteenth-century physics and the resulting crudely reductionist philosophy.[4]" In other words, Koestler and colleagues had had enough of mind being divorced from matter. What came out of the meeting—according to Kurt Stange, editor of *Annals of Family Medicine*—was "a chain of evidence that biological and social phenomena, like molecular and physical occurrences, evolve as events with many degrees of freedom, but with 'ordering restraints exerted upon them by the integral activity of the 'whole'.[5]"

A year prior to the convening, Koestler published the book "The Ghost in the Machine" in which he introduced the concept of a holon, described as a "whole part"[6]. Nothing is separate and nothing is together. Everything is contextualized, nested. Everything flows into everything else, then exits the same way it enters (See Figure 1). An atom is part of a molecule, a molecule part of a cell, a cell part of an organ, an organ part of a person, a person part of a household, a household part of a neighborhood, a neighborhood part of a city, a city part of a state which, in turn, is part of a nation-state which, in turn, is part of the geopolitical body politic. Higher up, policy decisions are made. These decisions impact many, doing so across geographic scales that hierarchically drift from international edicts to federal laws to state and local ordinances to neighborhood and household conditions, where it all ultimately lands into the geography of the body, including the well-being of a birth mother and their yet-born child.

In retrospect, what Koestler was driving at is what's been subsequently termed the "social determinants of health" (SDOH): a concept that's moved from the borders of academia into the discourse of a general consensus. The shorthand explanation of SDOH is that 80% of one's health happens outside of the doctor's office, with the other 20% in the hands of healthcare practitioners[7]. Put another way, issues like food, housing, schooling, smoking, stress, etc. matter. Consequently, SDOH has become a hot-button topic that's proving to be a looking glass for the healthcare industry specifically, and for the American experiment generally—particularly that balance between economic growth and societal progress.

This need for self-reflection isn't new. Our national accounting system, or Gross Domestic Product (GDP) counts things like imprisonment, air pollution, war production, cigarette ads, gun purchases, explained Robert Kennedy in 1968. Yet it does not allow for "the health of our children", "the strength of our marriages",

Figure 1 Systems of hierarchy. Source: Engels, G. 1980

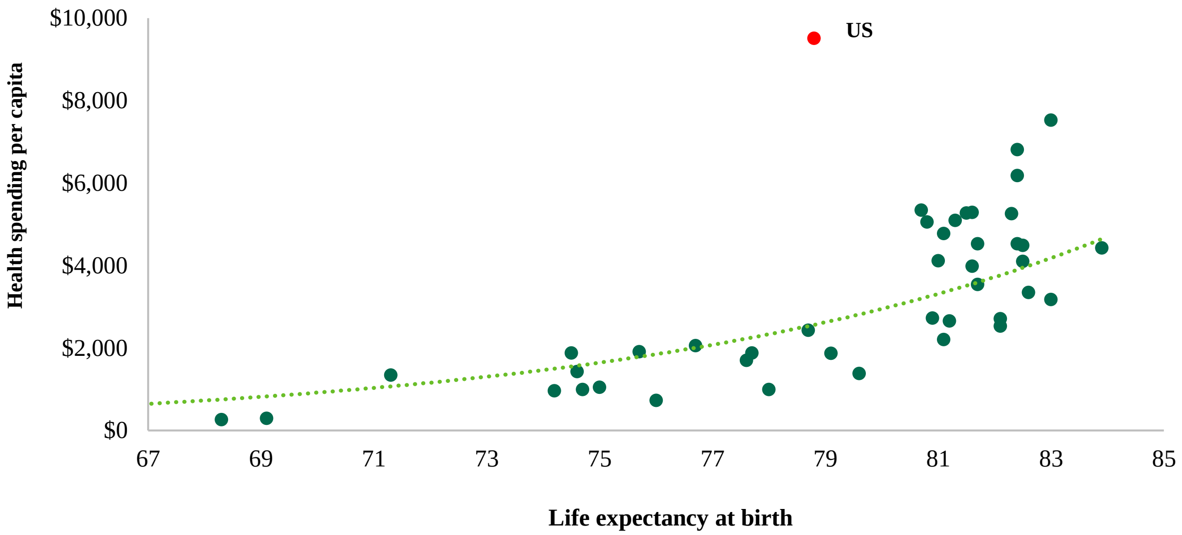


[4] Koestler A, Smythies JR, eds. Beyond Reductionism: New Perspectives on the Life Sciences. Boston, MA: Houghton Mifflin Co; 1971.
[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2746514/#r4
[6] https://journals.sagepub.com/doi/pdf/10.1177/00333549141291S206
https://www.researchgate.net/publication/270338868_The_Holonic_Revolution_Holons_Holarchies_and_Holonic_Networks_The_Ghost_in_the_Production_Machine
[7]

"the intelligence of our public debate", Kennedy continued. "It measures everything…except that which makes life worthwhile.[8]"

Figure 2 is illustrative. It shows the U.S. has the highest per capita spend on healthcare services compared to peer nations, but its life expectancy lags. That is, what's good for GDP does not always relate to what's good for individuals.

Figure 2: Life expectancy versus health spending per capita. Source: OECD, 2018.



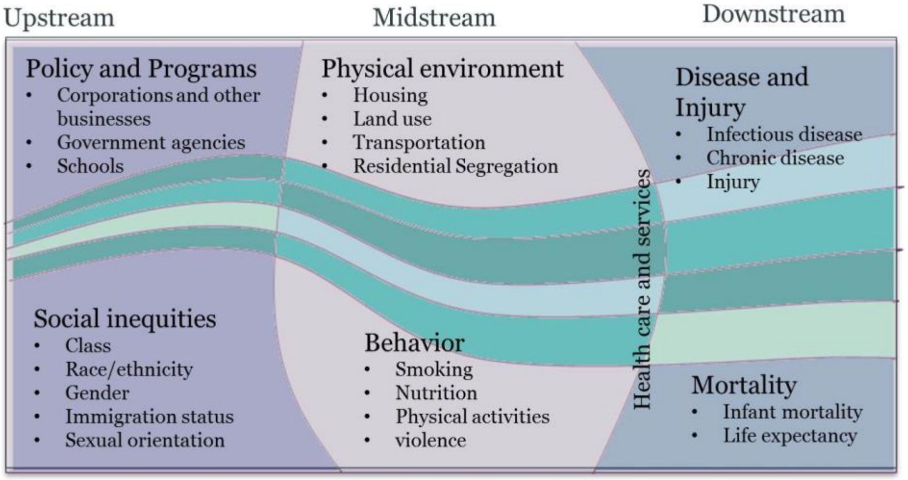
With disabusing stats like these, the arrival of SDOH onto the mainstream of public policy is upon us. "The Biden administration is expected to fortify existing policies and increase funding to deal with social determinants of health," explains *Inside Policy* Hea*lth*[9]. To the extent the field is mature enough to effectively capitalize on this investment is uncertain. Regardless, the field is ripe for discovery, given that some basic questions still need explored. Namely, where do social determinants start? Where do the bodily effects end? As for the former question, the field may have gotten off to a bit of a false start. A person's zip code, for instance, has come to stand in for the key spatial area of impact. "Zip code better predictor of health than genetic code," reads a Harvard School of Public Health analysis[10]. Echoes a Pub Med analysis, "Why Your ZIP Code Matters More Than Your Genetic Code: Promoting Healthy Outcomes from Mother to Child.[11]"

The problem, here, is that zip codes do not represent an actual area on a map, but rather a collection of routes that help postal workers deliver mail. As such, they aren't designed to measure demographic and socioeconomic trends. "They are not a good representation of actual human behavior," explains one geospatial analyst, "and when used in data analysis, often mask real, underlying insights, and may ultimately lead to bad outcomes.[12]" A better alternative is the use of census tracts, or areas equivalent to a neighborhood. Census tracts are established by the Census Bureau for analyzing populations between 2,500 to 8,000 people.

[8] https://www.theguardian.com/news/datablog/2012/may/24/robert-kennedy-gdp
[9] https://insidehealthpolicy.com/daily-news/biden-hhs-expected-beef-social-determinants-health-efforts
[10] https://www.hsph.harvard.edu/news/features/zip-code-better-predictor-of-health-than-genetic-code/
[11] https://pubmed.ncbi.nlm.nih.gov/27513279/
[12] https://carto.com/blog/zip-codes-spatial-analysis/

Another issue arises in the conceptual framing of how social determinants come about. The preference is to focus on **midstream** factors, or the interplay between individual- and neighborhood-level indicators (See Figure 3). Things like access to parks and grocery stores as place-based factors, or smoking and diet as an individual, behavioral-based factor. No doubt, neighborhood- and individual-level factors are integral in understanding how one's immediacy affects one's longevity. Dietary habits, for instance, account for numerous **downstream** effects, like diabetes. Couple that fact with the understanding that what one eats is associated with the quality of food accessible in one's neighborhood, then it makes sense that the geography of intervention is at the level of the community. Think food pantries, community gardens, farmer's markets, etc. That said, just as a household is downstream from a neighborhood, so is a neighborhood downstream from a region, and a region from a nation. Elaborating, **upstream**, or structural, factors are key in how SDOH play out, be it how macroeconomic policy has affected regional labor markets (e.g., privatization, deindustrialization), or how institutional racism has affected a neighborhood's housing market (e.g., redlining). "As the field of SDOH grows," writes the authors of the RAND paper "Understanding the Upstream Social Determinants of Health", "there is increasing emphasis on understanding and addressing the fundamental causes of poor health and inequities.[13]". Advancing the field of SDOH, then, means taking these upstream factors into account. This analysis addresses that need.

Figure 3



Part 1: Cleveland, Great Healthcare, Poor Population Health—A Descriptive Analysis

In the field of city building, numerous outcomes are tracked by policymakers and strategists to chart progress, such as jobs, income, housing price, poverty rate, etc. There is, however, arguably no better measure of a city's success than the life expectancy of its residents. If conditions in a city are poor, it will end up in the physiology of the people living there. Likewise, if things are good, well-being and longevity are the result.

How does Cleveland and Ohio fair in this regard? Less than ideal. According to stats from the CDC, Ohio's life expectancy of 76.8 ranks 12th worst in the nation (See Figure 4). Figure 5 maps life expectancy by county for the State of Ohio. Cuyahoga ranks 41st out of 89 counties. This, despite not only having the best healthcare system in Ohio, but also one of the tops in the country—which again speaks to the disunion between economic and societal development; or healthcare as a global industry versus healthcare as part of a menu of local assets to better population health.

[13] https://www.rand.org/content/dam/rand/pubs/working_papers/WR1000/WR1096/RAND_WR1096.pdf

Figure 4: Life Expectancy by State. Source: National Center for Health Statistics, 2018.

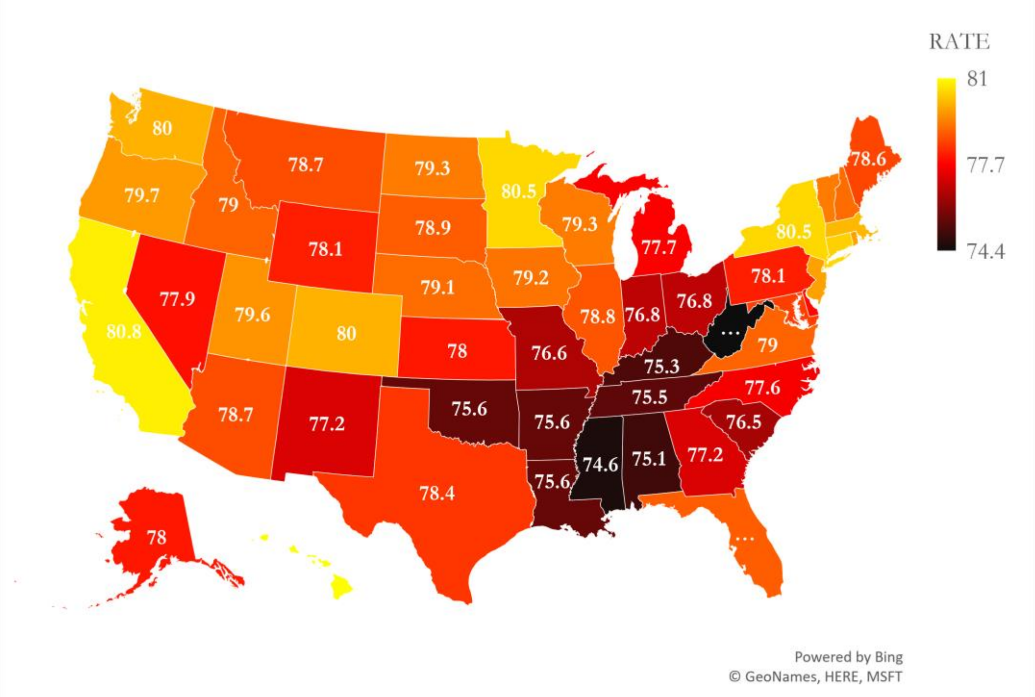


Figure 5: Life Expectancy for Counties in Cuyahoga County. Source: County Health Rankings, 2020.

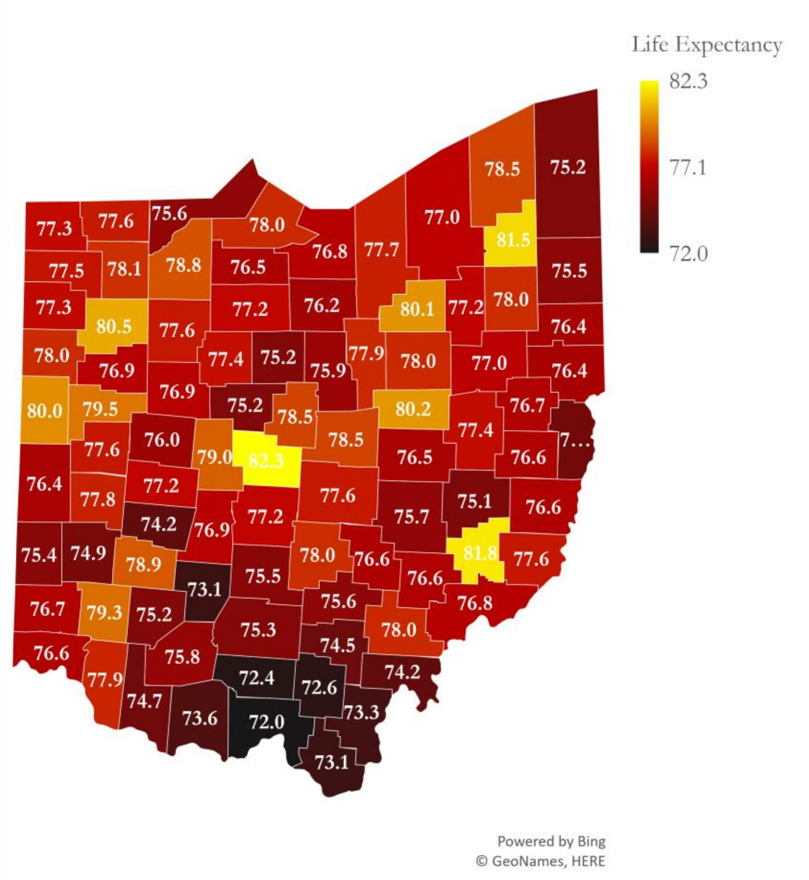


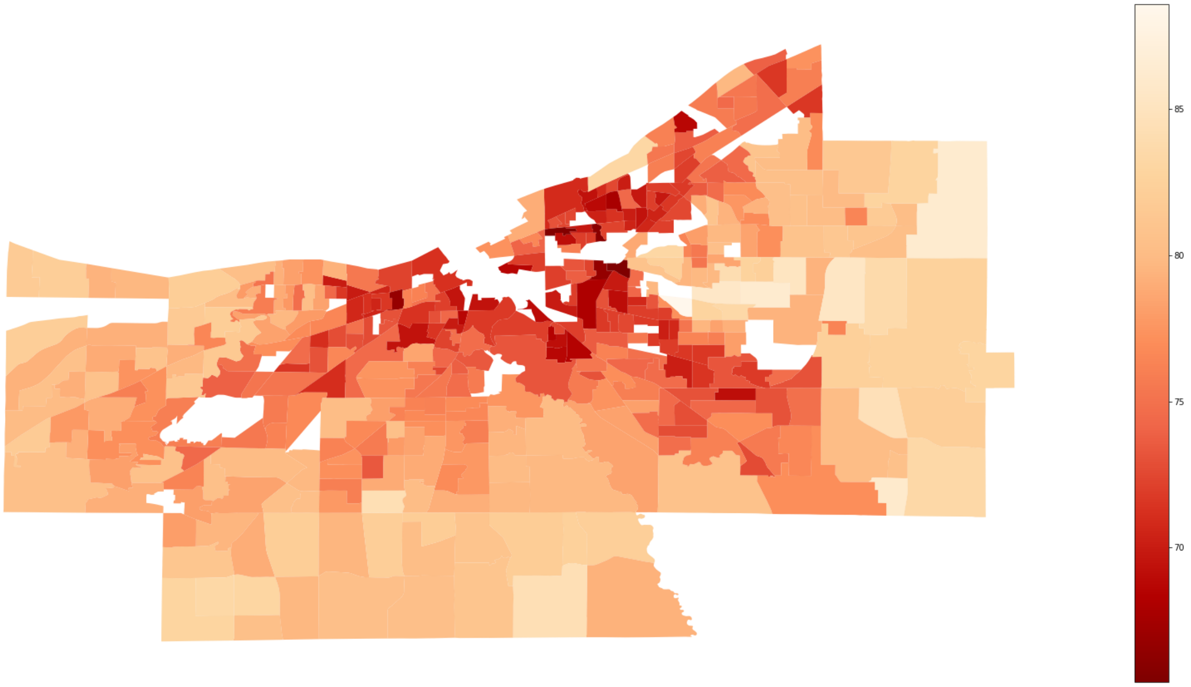
Table 1 shows that Cuyahoga County's life expectancy figures are racially disparate. On average, Black Cuyahoga County residents live to an age of 73.6, below that of White (78.3), Asian (89), and Hispanic or Latino (82.7) residents. This racial disparity is present in all counties within the Cleveland metro (See Table 1).

Table 1: Life Expectancy for Counties in Cleveland's Metropolitan Statistical Areas. Source: County Health Rankings, 2020.

	Life Expectancy	Life Expectancy (Asian)	Life Expectancy (Black)	Life Expectancy (Hispanic)	Life Expectancy (White)
Cuyahoga	77.0	89.0	73.6	82.7	78.3
Lorain	77.7	88.5	72.3	81.2	78.0
Lake	78.5	91.1	75.6	90.6	78.3
Medina	80.1	90.4	76.5	94.4	80.0
Geauga	81.5	N/A	N/A	N/A	N/A

This disparity manifsts at the geography of the neighborhood. Figure 6 shows life expectancy for all Cuyahoga County residents at the the census tract level. Higher life expectancies are found in the outter suburban areas of the county; whereas lower life expectancies are in the city proper of Cleveland, particularly in Cleveland's historically-Black Near East Side neighborhoods.

Figure 6 Life Expectancy in Cuyahoga County. Source: U.S. Small-area Life Expectancy Estimates Project-USALEEP, 2010-2015.

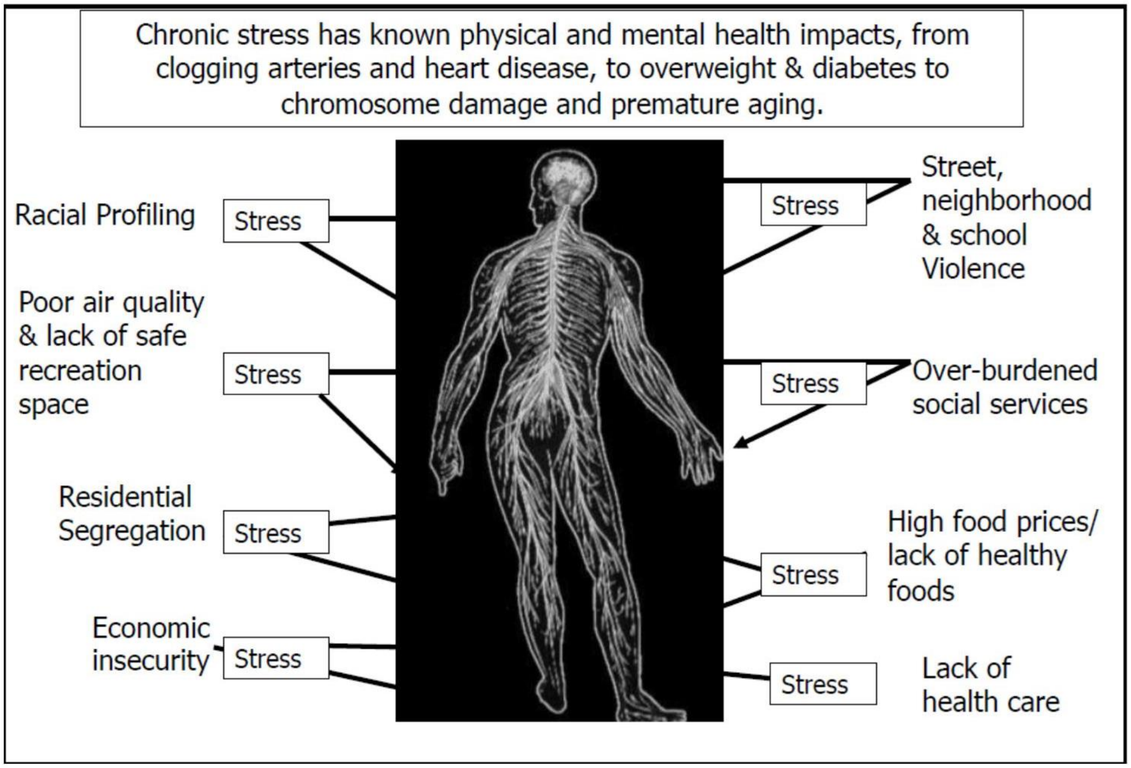


That Blacks have lower life expectancies in Cleveland is not a revelatory finding. There's been a substantial body of research documenting racial disparities in life expectancy[14]. Cal Berkeley's Jason Corburn, for instance, explains that the wear and tear of chronic, psychosocial stress—what Baratunde Thurston refers to as "living while black[15]"—creates for a persistently alerted nervous system, and the fact that the "engine is always running" has a bodily effect (See Figure 7). This effect is incurred through the process of epigenetics,

[14] https://www.theatlantic.com/ideas/archive/2019/10/too-short-lives-black-men/600628/
[15] https://www.baratunde.com/livingwhileblack

which is an emerging field examining how the "outside" affects the "inside" (and vice versa). "Social epigenomics is defined as the study of how social experiences affect our genes and biology," explain the authors of "Understanding the Interplay Between Health Disparities and Epigenomics[16]". Though nascent, the field is sure to disrupt the outdated mind/body dualism Koestler and his ilk were crowing about. Moreover, it's a type of holism desperately needed going forward. "Social epigenomics is uniquely positioned at the intersection of population health and precision medicine," explains one epigeneticist, "allowing us to understand how exposure to social and environmental stressors modifies the way in which genes are expressed and ultimately alter our risk for disease."

Figure 7 Source: Adapted from Professor Jason Corburn, University of California, Berkeley


Chronic stress has known physical and mental health impacts, from clogging arteries and heart disease, to overweight & diabetes to chromosome damage and premature aging.
Racial Profiling
Stress
Poor air quality & lack of safe recreation space
Stress
Residential Segregation
Stress
Economic insecurity
Stress
Street, neighborhood & school Violence
Stress
Over-burdened social services
Stress
High food prices/ lack of healthy foods
Stress
Lack of health care
Stress

In the case of expecting mothers, the impacts of SDOH are twofold, affecting the health conditions of the mother, as well as the fetus (pre-birth) or infant (post-birth). As life expectancy is a critical indicator gauging the success of a place, a sub-indicator of life expectancy, or infant mortality[17], is perhaps the measure of all measures. Infant mortality has long been viewed "as a synoptic indicator of the health and social condition of a population," so notes the writers of the "The First Injustice"[18], with the term "synoptic" having a Latin lineage referencing an "accounting of the times"[19]. Worldwide, rates of infant mortality have decreased. Figure 7 details OECD data for the G 20 countries. The U.S., for instance, went from an infant mortality rate of 26 in 1960 to 5.8 in 2017. Within the U.S., infant morality rates vary. Ohio's infant mortality rate, 6.97, ranks 9th worst in the nation, tied with South Carolina (See Figure 9). It's also lowest in the Midwest, below that of its peers Pennsylvania (5.85), Michigan (6.33), and Illinois (5.52).

[16] https://www.frontiersin.org/articles/10.3389/fgene.2020.00903/full
[17] Note: Infant mortality is defined as the number of deaths of those under the age of 1 per 1,000 live births.
[18] https://www.jstor.org/stable/2952547?seq=1
[19] https://en.wikipedia.org/wiki/Synoptic_Gospels

Figure 8: Infant Mortality (deaths per 1,000 births) for G20 Nations. Source: OECD (USA in red).

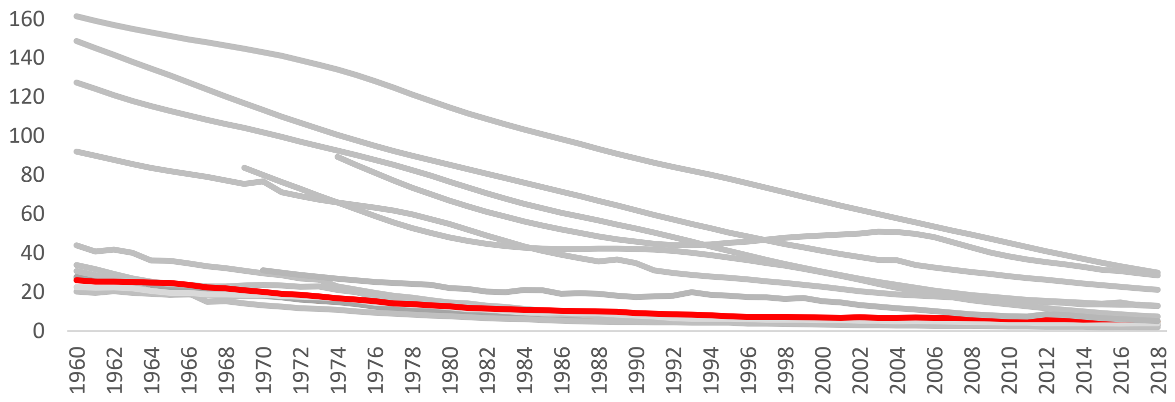


Figure 9: Infant Mortality by State. Source: National Center for Health Statistics, 2019.

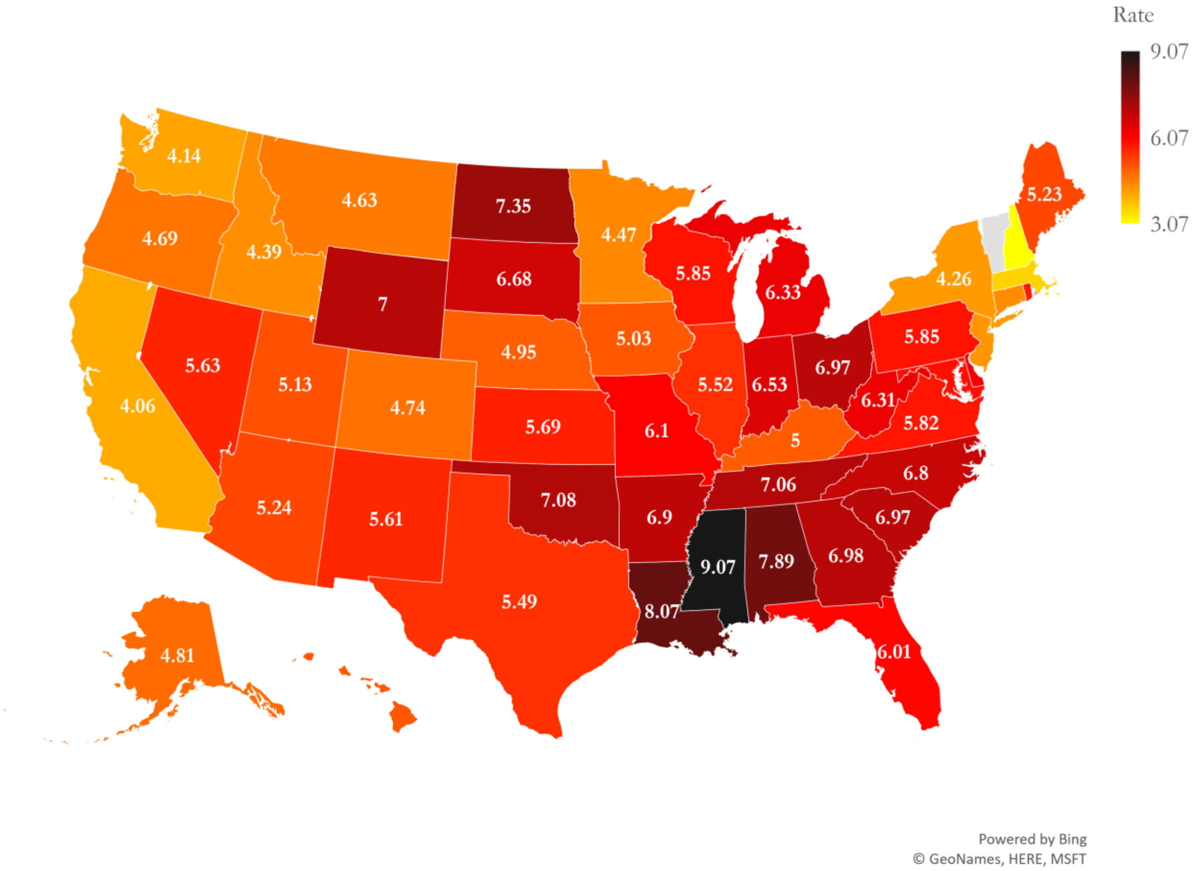


Figure 10 shows Ohio's infant mortality rates vary dramatically by race, with Black Ohioans (14.3) having more than double the rates of Whites (5.1) and Hispanic and Latinos (5.8). Figure 11 shows the breakdown of infant mortality rate by county. Cuyahoga performs poorly, with only the Appalachian county of Guernsey performing worse (10).

Figure 10: Infant Mortality Rates by Race and Ethnicity in Ohio. Source: Ohio Department of Public Health

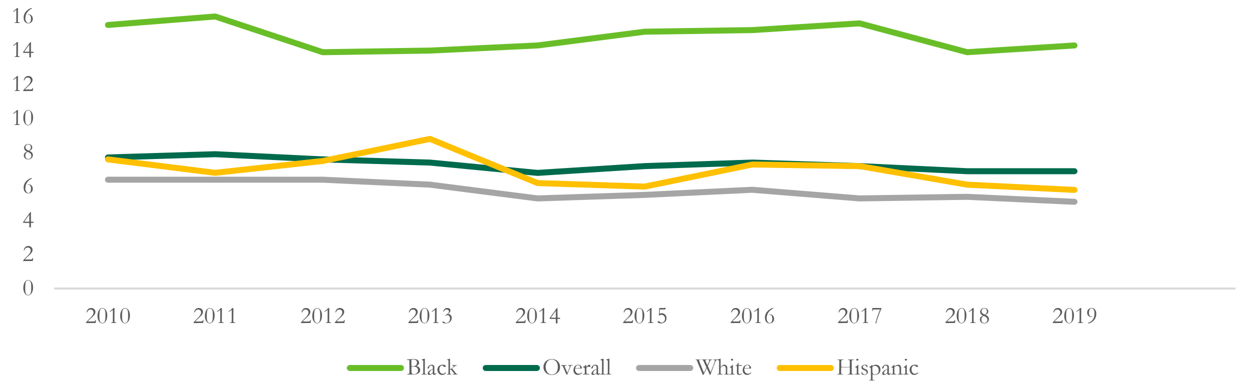
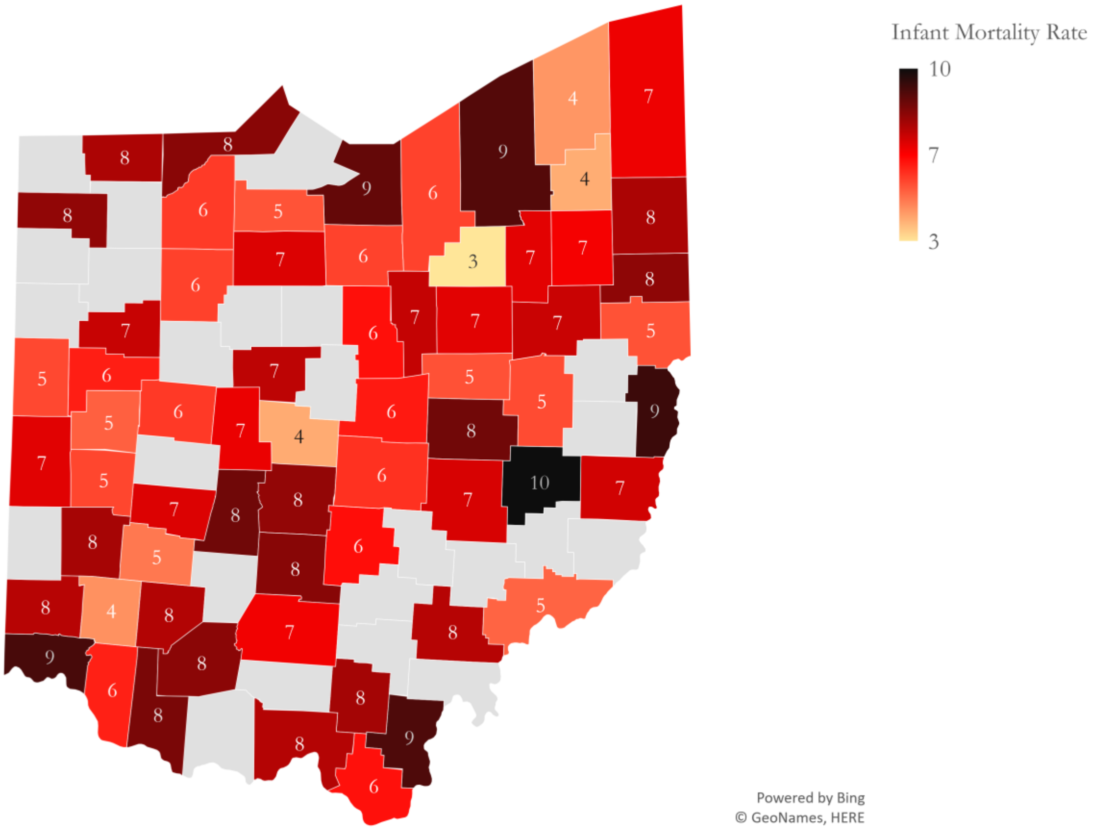
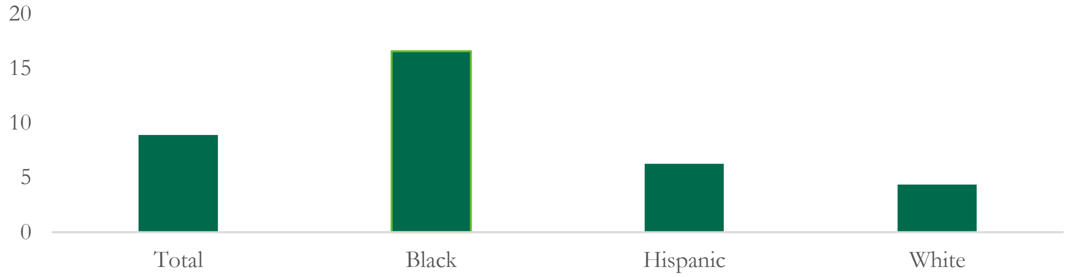


Figure 11: Infant Mortality Rate by County. Source: 2021 County Health Rankings, data from 2015-2019



Cuyahoga's higher infant mortality rate is driven by racial disparities (See Figure 12). The infant mortality rate for Blacks is quadruple that of Whites (4) and nearly triple that of Hispanics (6).

Figure 12: Infant Mortality Rate Cuyahoga County. Source: 2021 County Health Rankings, data from 2015-2019.



None of the above stats, however, are telling of anything that we didn't already know, i.e., that race is associated with numerous poor outcomes, be it low life expectancy or high infant mortality rates. Matter of fact, so much of social science research can get passed off as informative when it's really just confirmative. Showing racial disparities amidst various economic, socioeconomic, and well-being outcomes is arguably at the tops of this list. An unpacking of these descriptive analyses is thus needed. We need awareness beyond the fact that race matters. Rather, how does it matter? In the case of this analysis, that entails attempting to show how racial disparities in infant mortality and life expectancy outcomes come about. What is the package of social determinant features that best explain the difference? How long-shadowed are these features? Do they rest in the spatial realm of individual behavior where intervention is more direct? Think smoking cessation or pre-natal checkups. Or are there features that are cast further back, indicative of long-term, structural, factors that have settled in to create for an overall setting of psychosocial hardship that gets manifest in everyday thoughts, feelings, and acts—and ultimately individual well-being?

Explains Cleveland-based Christine Farmer, former CEO of Birthing Beautiful Communities in an interview she did with *Shoppe Black* entitled "This Doula Is Fighting Back Against the High Infant Mortality Rate in her Community": "If we want to see our babies staying alive, our mothers healthy, our men as protectors and providers, we have to get to the root and address these problems.[20]" Or as one team of population health researchers put it: "It's time to consider the causes of causes.[21]"

[20] https://shoppeblack.us/2018/09/doula-fighting-infant-mortality/
[21] https://journals.sagepub.com/doi/pdf/10.1177/00333549141291S206

Part 2: Swimming Upstream, An Explanatory Analysis

Modern technologies, particularly the availability of big data, affordable data storage, and computational processing that is cheaper and more powerful, has expanded the application of data science. Yet while the capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics, the advent of data science in social science has been much slower[22]. This is unfortunate. Social science by its nature is about the study of complex systems, and patterning outcomes like infant mortality and life expectancy is an undertaking that borders on the impossible given the seemingly endless factors that need considered. That said, the aforementioned field of epigenomics is "fool's gold" unless the same computational rigor that maps humans cellularly can be used to map humans societally[23]. We can't do precision policy[24] unless we know what levers to pull and why. "Against a background of financial crises, riots and international epidemics," explains one group of computational social scientists, "the urgent need for a greater comprehension of the complexity of our interconnected global society and an ability to apply such insights in policy decisions is clear.[25]" This analysis is intended to be a step in that direction.

The Dataset

We created a social determinant of health (SDOH) data set at the census tract level that pooled information from multiple sources. Millions of variables were architected across tens of thousands of census tracts nationwide. The sources of those variables include:

- American Community Survey 5-Year Estimates and Decennial Census, 1970 to 2019. The data contains various economic, socioeconomic, demographic, household, and housing data. It is available for each tract in the United States.
- Migration Database. This proprietary database examines migration trends for every census tract in the U.S. by looking at 10-year age cohort changes for the decades 1970 to 2017. Data is from the Census, and it shows what communities are losing or gaining people by age groups across a ten-year period, with the notion that migration is a leading indicator to investment.
- Longitudinal Household Employment Dynamics, or (LODES), which breaks down employment by NAICS industry for each employee in a census tract, or where the jobs are, as well as every working resident in a census tract, or where the workers are. Worker and resident characteristics by demographics are also available, as are commuting patterns.
- Employment Sector. This proprietary database breaks down NAICS industry data into three sector types: Knowledge Workers, Lower-Wage Service Workers, and Manufacturing Workers using LODES data for each tract in the U.S. The rationale for this dataset relates to the fact of a labor market bifurcation between knowledge and service workers that is manifest as "residential sorting", or the settlement of workers by class features. This sorting is a leading indicator of investment and thus dictates social determinant patterns.
- Mortgage data (HMDA) charts mortgage applications, rejections, and approvals by demographic and place at the census tract level. These data shed light on lending patterns that could be discriminatory. It is the most comprehensive source of publicly available information on the U.S. mortgage market.
- NASA's Carbon Monitoring System (CMS) data, which includes CO2 emissions monitoring data.

[22] https://dash.harvard.edu/bitstream/handle/1/4142693/King_Computational.pdf;jsessionid=9B98546568B052E6AAE3FC6A23EA5B36?sequence=2

[23] https://cebp.aacrjournals.org/content/cebp/22/4/485.full.pdf

[24] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6761592/

[25] https://ui.adsabs.harvard.edu/abs/2012EPJST.214..325C/abstract

- Northeast Ohio Community and Neighborhood Data for Organizing (NEOCANDO) property data, which includes data on property characteristics, housing sales, foreclosures, and sheriff's sales.
- the Environmental Protection Agency's (EPA) Smart Location Database[26] that is designed to measure "location efficiency". It includes more than 90 categories [27] including employment, job diversity, demographics, housing density, diversity of land use, neighborhood design, destination accessibility, and transit service.
- The EPA's Environmental Justice Screening and Mapping Tool (EJSCREEN)[28] which includes demographic and environmental indicators, as well as EJ indexes (which is a combination of the two). Environmental indicators include, lead paint risk, air toxicity cancer risk, respiratory hazard risk, particulate matter levels, ozone levels, proximity to traffic and congestion, proximity to wastewater discharge, and proximity to hazardous waste sites.
- Yelp, "The Dataset"[29], includes a list of businesses, schools, hospitals, and other government- for-profit- and non-profit institutional assets for the Cleveland metro, including their business attributes. E.g., hours, parking, multi-modal availability, ratings, and check-ins.
- CDC's 500 Cities Project 2017[30], which includes estimates for chronic disease risk factors, health behaviors, health outcomes, and clinical preventive services use.
- Small-area Life Expectancy Estimates Project-USALEEP provides valid life expectancy estimates the level of the census tract.
- CDC National Vital Statistics System (NVSS) The linked birth and infant death data set was comprised to illuminate the complex inter-relationships between infant death and risk factors present at birth. The lowest level of geography the data is available is at the county level. Information from the death certificate is linked to the information from the birth certificate for each infant under 1 year of age who dies in the United States. Deidentified variables include age, race, nativity, and Hispanic origin, birth weight, period of gestation, prenatal care usage, maternal education, live birth order, marital status, and maternal smoking, adverse birth events, among others[31].

The Model(s) and Results

The most fundamental question one can ask of a model is what features have the biggest impact on predictions. This is called "feature importance"[32] and it is based on the idea that more important features have more predictive impact. Practically speaking, there are two reasons why feature selection is used[33]. First, to reduce the number of features and thus the likelihood of overfitting the model (believing a feature is important when it isn't.). Second, to improve the generalization of models. Simply, we want to make the model interpretable and actionable, while still embedding it with complexity, or a vastness of variables, as is required by the reality being modeled.

The algorithm chosen for the feature selection is called Random Forest[34]. How does this algorithm know which features have the highest predictive power? We can look at the problem from a different perspective. If a feature is important, then if it's removed from a model the accuracy of that model would decrease. Based on a series of decision trees, Random Forest "loops" through pre-identified columns in a

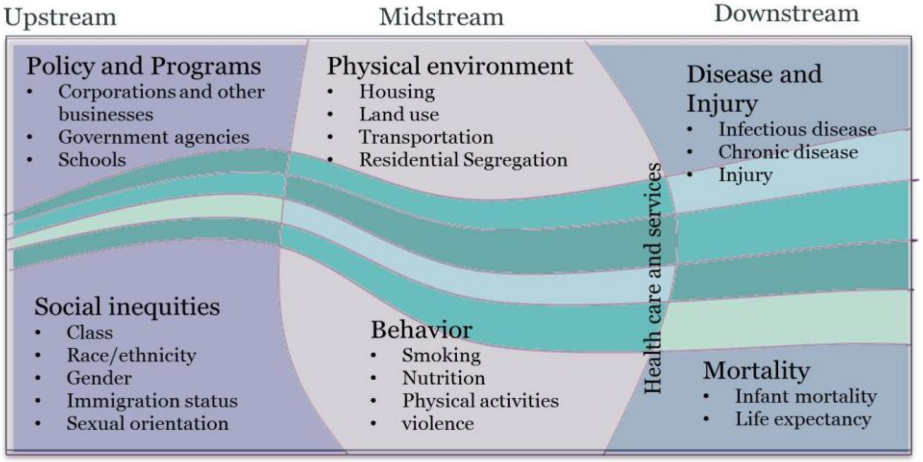
[26] https://www.epa.gov/smartgrowth/smart-location-mapping#SLD
[27] https://www.epa.gov/sites/production/files/2014-03/documents/sld_userguide.pdf
[28] https://www.epa.gov/ejscreen/what-ejscreen
[29] https://www.yelp.com/dataset
[30] https://www.cdc.gov/places/
[31] https://www.cdc.gov/nchs/nvss/linked-birth.htm
[32] https://towardsdatascience.com/explain-your-machine-learning-with-feature-importance-774cd72abe
[33] https://blog.datadive.net/selecting-good-features-part-iii-random-forests/

[34] https://medium.com/@amaarora/implementing-feature-importance-in-random-forests-from-scratch-2216e031ff74

dataset while making predictions. While doing so, it shuffles the column and makes predictions with the shuffled column. If a column is important to making predictions, shuffling it should lead to an increase in the error term. Therefore, those columns which lead to a maximum increase in loss function are most important Interpreting a Random Forest model is relatively simple: variables are ranked by the magnitude of their loss functions.

Two Random Forest models were built to answer two questions. The first question asks: **What SDOH features predict risk for infant mortality?** Variables can be categorized as upstream, midstream, and downstream (See Figure 13). The outcome variable was the occurrence of infant death via the NVSS linked birth and infant death data set. The explanatory variables assessed various features of the birth mother, the birth father, the infant, and the birth event. Data were available for all infant deaths between 1995 to 2017.

Figure 13



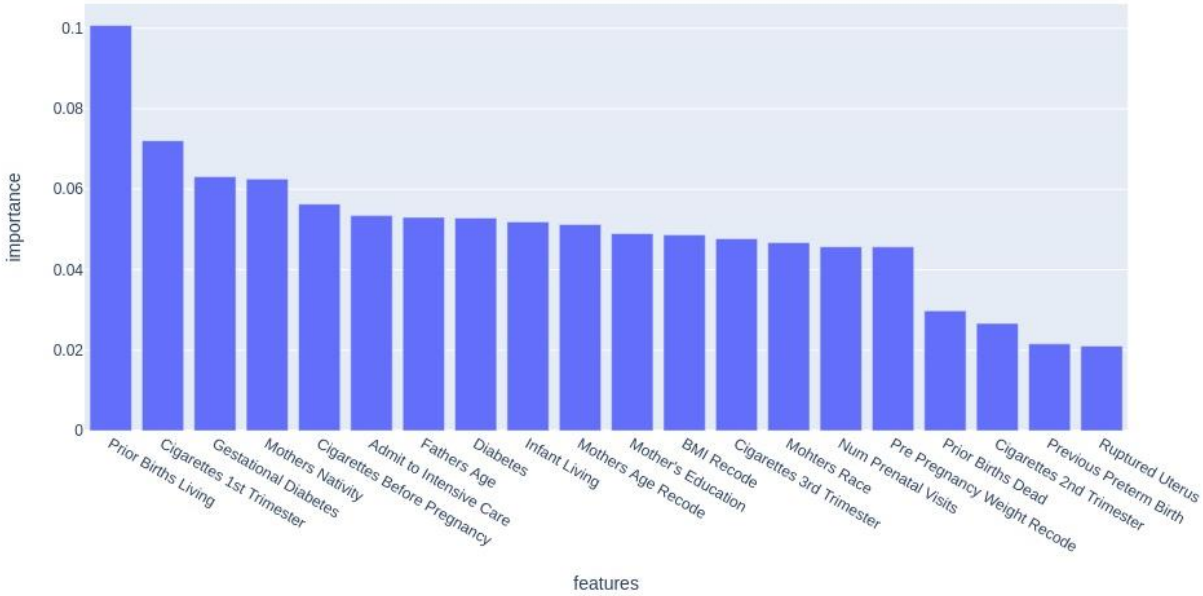
Notably, this type of data is administrative[35]. It is not estimated but documents actual occurrence, with each incidence aggregated "up" to the county level for the nation as a whole. This does not mean the geography of interest from a SDOH standpoint is the county. A county is too large to discern the spatial gradience inherent in SDOH. Rather, the "geography" of interest, here, is roughly equivalent to the household (e.g., mother, father, and infant characteristics). More acutely, there's the care dyad that is pregnancy. Arguably, the most direct line that can be drawn to the health of a newborn is via the health of the birth mother. From a spatial standpoint, one can conceive of the maternal-fetal dyad[36] as one "geography" that's nested together. What influences one strongly influences the other.

It comes as no surprise, then, that individual-level, downstream and midstream factors relating to the birth mother figure heavily in the results (See Figure 14). This includes the presence of comorbid conditions. Gestational diabetes was the 4th highest risk factor for infant death, along with diabetes (7th), an elevated BMI (11th), and pre-pregnancy weight (16th). These conditions, in turn, are related to behavioral midstream factors of the mother that proved to have predictive power. Cigarette smoking in the 1st trimester was the 2nd highest risk factor, with pre-pregnancy smoking 5th, smoking in the 3rd trimester 13th, and smoking in the second trimester 19th. Another behavior that proved important was the number of prenatal visits to healthcare providers (14th.) Other factors that predicted infant mortality were birth-related events (admission to NICU [9th highest risk factor] and a ruptured uterus during delivery [18th highest risk factor]. Lastly, a "past-as-prologue"[37] set of features emerged. The highest risk in our model was the mother's prior births living, with prior births dead (17th) and previous pre-term birth (20th). These latter factors speak to the importance of the experiential-driven "doula" movement as an intervention in infant mortality, one that is highly espoused by Birthing

[35] https://www.sciencedirect.com/science/article/pii/S0049089X1630206X?via%3Dihub
[36] https://www.jstor.org/stable/3562716
[37] https://pubmed.ncbi.nlm.nih.gov/30539421/

Beautiful Communities in Cleveland. From an intervention standpoint, then, the results lend themselves to an individual- behavioral-based approaches that are standard in prenatal care. These approaches foster the health of the mother via programs like smoking cessation and nutrition promotion, as well as the nurturing of fetal development through prenatal visits—all actionable, proven interventions.

Figure 14: Random Forest Model for Infant Mortality Risk. Source: Author's Calculations



Yet such interventions are not "swimming upstream", or tackling root causes. There's substantial research, for instance, showing that psychosocial stress is a precursor to health behaviors[38], particularly in communities of color wherein stressors are continuous and multi-faceted[39]. "Robust associations exist between current tobacco smoking and perceived stress," note the authors of "Racial/Ethnic Differences in Physiological Stress and Relapse among Treatment Seeking Tobacco Smokers," "such that smokers report greater stress compared to nonsmokers. Multiple types of stress, such as finances, relationship conflicts, and discrimination are positively related to persistent smoking and inversely associated with cessation".

This helps explain why mothers race was 12th highest risk factor in our model. It's an upstream, structural factor—one in which other factors, like health behaviors, "filter" through. Not unrelated is the nativity of the mother (3rd highest risk factor), or foreign-born versus native-born, with immigrants showing lower risk in past research[40]. Other socioeconomic and demographic indicators also proved predictive including the mothers' education (16th highest risk factor), the mother's age (10th) and the age of the father (6th).

Taken together, infant mortality interventions are inherently hamstrung if geared only to the behavior of the birth mother. Simply, a mother's health and infant's death are not disconnected from the world around them. Well-being does not grow in a vacuum. It's influenced by the home the mother lives in, the street she lives on, and the city she resides in. Such local conditions, however, are carved out by the "hands" of structural policy. These federal and state policies could be economic, institutional, sociological, etc.

[38] https://www.researchgate.net/publication/221864259_Psychosocial_Stressors_and_Cigarette_Smoking_Among_African_American_Adults_in_Midlife

[39] https://journals.sagepub.com/doi/abs/10.1177/1054773817713420

[40] https://pubmed.ncbi.nlm.nih.gov/1994686/

Which leads us to the second model in the analysis that answers the question: **What upstream and midstream variables impact life expectancy?** Here, the outcome variable was life expectancy at the census tract level. The model was restricted to the census tracts within Cuyahoga County. The explanatory variables included numerous factors made available in our SDOH dataset. (Note: while infant mortality and life expectancy are strongly correlated, they aren't substitutes. Nonetheless, a robust life expectancy model will allow us to discern insights on structural SDOH that can have crossover impact on a number of downstream outcomes, including infant mortality.)

Before going further, the two maps below will serve to orient the reader. The first map is life expectancy in Cuyahoga County. The darker the red the lower the life expectancy. The map on the right is the percentage of workers who work in lower-wage service work, e.g., retail, food and accommodations. The darker the purple, the higher concentration of lower-wage service workers. The more yellow, the higher the concentration of knowledge workers, i.e., the "creative class". This is based on where they live, not where they work. Note the spatial overlap: Where lower-wage workers live there's less longevity. This is a concrete example of upstream factors—i.e., work and wages—being associated with a downstream health effect, i.e., dying. This is supported by Figure 16 showing a strong positive correlation ($r^2 = .78$) between per capita income and longevity in Cuyahoga County's census tracts.

Figure 15: Life expectancy (left), Source: Project USA ALEEP. Where Lower-Wage Workers Live (right) , Source: LODES.

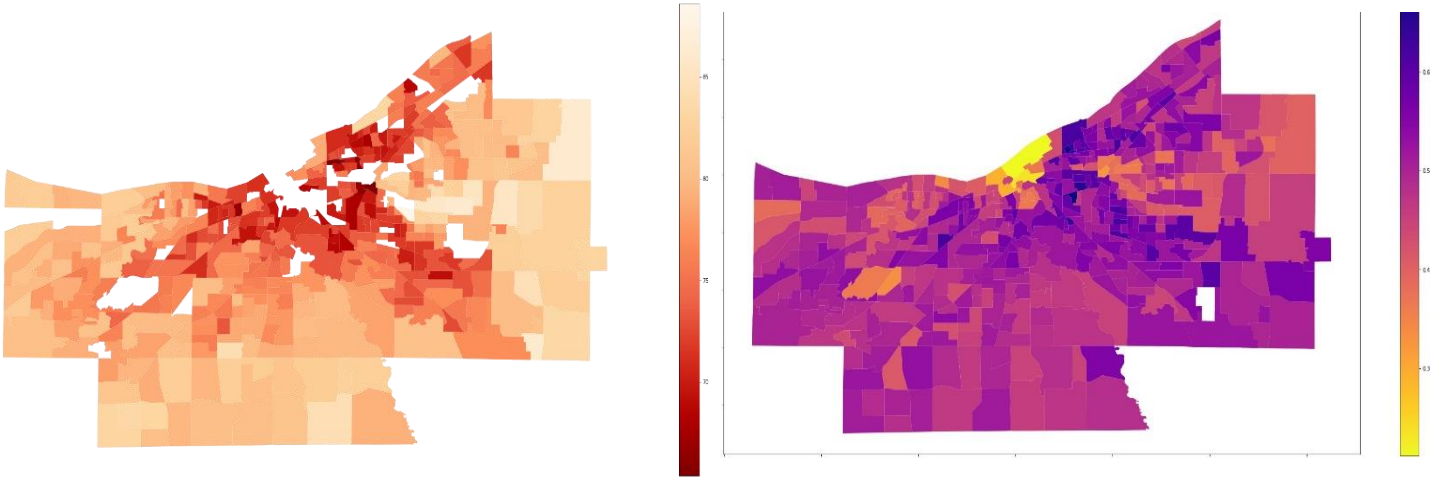
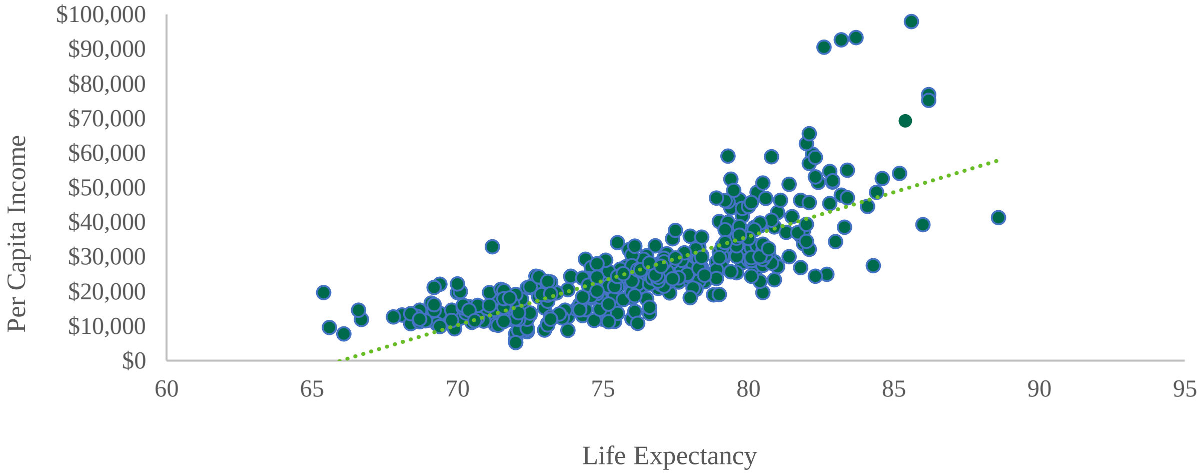


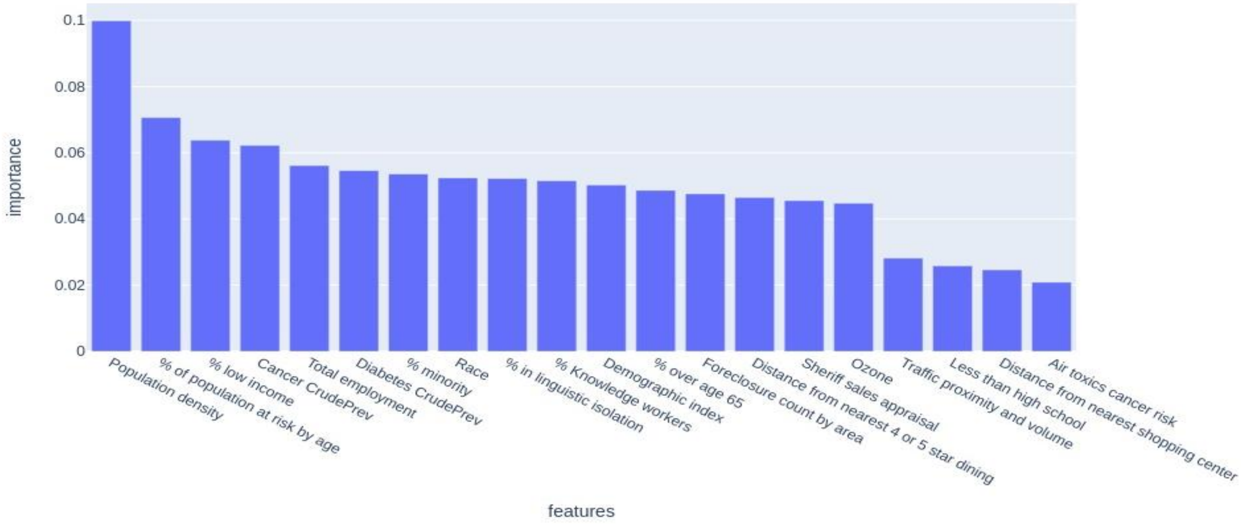
Figure 16 Life Expectancy v. Per Capita Income for Census Tracts in Cuyahoga County. Source: U.S.USALEEP, 2010-2015 and ACS 5-Year,2015.



Still, how do we get from work to wages to life and death? After all, there's lots of dots to connect in between. Past work by the current authors have detailed how macroeconomic policies such as deregulation, privation, and financialization have favored capital over labor, leading to a bifurcated labor market between knowledge workers on one hand and service workers on the other (with an eroding blue-collar sector in between) [41]. This bifurcation at the scale of the regional labor market is manifest as a bifurcation of the local housing market, a phenomenon called "residential sorting"[42]. The sorting gets played out in the way amenities flow. Knowledge- and tech-worker neighborhoods are flush with investment, manifest as a cornucopia of goods and services that check-off Maslow's hierarchy of needs: physical safety, healthy food, clean air and water, quality housing, good schools and healthcare, pretty aesthetics and parks, a strong social fabric and concomitant information access, not to mention the freedom from scarcity that allows the luxury of aspiration. Meanwhile, disamenities grow in areas of isolation: violence and trauma, dirty air and water, deteriorating housing, poor schools and health services, a digital divide, a social bond break with less information and support, and a lack of a psychological reprieve that comes with existing without enough.

Does our algorithm provide empirical support for how this conceptual model plays out? Out of thousands of features modeled, the top 20 with the highest predictive power are in Figure 17. They will be referenced by the extent they are more upstream or downstream, not the magnitude of their loss function.

Figure 17: Random Forest Model for Life Expectancy. Source: Author's Calculations



Keeping with the theme of macroeconomic factors, the feature with the 5th highest predictive power for neighborhood life expectancy was total employment for neighborhood residents, or where the workers are. Economic opportunity via employment thus proved critical. But not all jobs are equal. The feature with the 10th highest predictive power of neighborhood life expectancy was the concentration of knowledge workers who reside in a neighborhood. Think of this feature as a proxy for gentrification which, in turn, is a process of neighborhood transformation that acts—via demographic change—as a signal to the market to invest.

Next are upstream features related to race and ethnicity. Race had the 8th highest predictive power, just behind the percent of residents who were minority, or those who were non-white (7th). Here, structural racism is key, particularly the long-standing link between racial bias and capital flow[43]. Relatedly, the percent

[41] http://thefutureofgrowth.com/fog/
[42] https://journals.sagepub.com/doi/full/10.1177/0042098018798759
[43] https://www.urban.org/features/structural-racism-america

of residents in linguistic isolation—defined as those living in a household in which all members aged 14 years and older speak a non-English language and also speak English less than "very well"[44]—predicted neighborhood life expectancy 9th best. Socioeconomically, the percent of residents that were low income had the 3rd highest predictive power, and the number of residents without a high school degree was 18th. A feature that combined both race/ethnicity and socioeconomic status was the Demographic Index. This feature was calculated as the average score of percent low-income and percent minority in a given neighborhood. It had the 11th highest predictive power.

We can keep going. Where residents reside or sort, so do features of a given housing and retail market. The amount of foreclosures and sheriff sales—two key indicators of housing distress—were the 13th and 15th most predictive of neighborhood life expectancy, respectively. When it comes to retail amenities, the proximity to Yelp-rated 4- or 5-star dining (another proxy for gentrification) was 14th, and the proximity to shopping centers was 19th.

Of course, the flipside of amenities are disamenities, or those things people do not want to live next to. When it comes to disamenities in SDOH, exposure to environmental toxins are a chief concern, if only because toxin-to body-to ill health is a rather straightforward set of chain of events. (See lead and Flint.) The environmental toxin features that proved important in our model included: ground-level ozone (16th), traffic proximity and traffic volume (17th) and air toxicity risk for cancer (20th). Each of these are conceptually related, and also tie in the highest risk factor in our model: population density. While density is in fact lauded as a preferential form of urban design when it comes to SDOH, if it is centered around the automobile (as it mostly is in Cleveland), then there's the propensity to make local health worse. To that end, all the the pre-mentioned upstream and midstream features flow downstream, eventually "landing" in the body as the presence of disease. Here, the prevalence of cancer in a community had the 4th highest predictive power on neighborhood life expectancy, and the prevalence of diabetes was 6th.

The Takeaway: Tilting Toward Sickness or Tilting Toward Health

In the book by geostrategist Thomas P.M. Barnett's called *The Pentagon's New Map*[45], he explains that the world is divided between two types of geographies: the "Core", where "globalization is thick with network connectivity, financial transactions, liberal media flows, and collective security," and the "Gap", or areas disconnected from globalization and defined by poverty, low education rates, low life expectancy, and "the chronic conflicts that incubate the next generation" of instability. While Barnett's "haves and have nots" was conceived at the level of the nation-state, it need not stay there. There is a Core and Gap between American regions, within regions, within cities, even within neighborhoods. "We ignore the Gap's existence at our own peril," concludes Barnett.

Not unlike Barnett's frame between Core and Gap, the takeaway, here, is that the social flow of opportunity and inopportunity from a SDOH standpoint is not equally dispersed. The topography is tilted, privileging not only individuals and their communities, but generations of individuals and their communities. Elaborating, structural factors globally manifest as inequalities locally, igniting psychosocial stress that changes the body's biology. And it's a sequence that lingers intergenerationally. "Each exposure has effects that may persist across the life course and in some instances may be transmitted to offspring via epigenetic inheritance," notes the authors of the essay "Biological memories of past environments: Epigenetic pathways to health disparities.[46]" "Since epigenetic markings provide a 'memory" of past experiences, minimizing future

[44] https://www.cdc.gov/pcd/issues/2006/jan/05_0055.htm

[45] https://www.esquire.com/news-politics/a1546/thomas-barnett-iraq-war-primer/

[46] https://www.researchgate.net/publication/51151168_Biological_Memories_of_Past_Environments_Epigenetic_Pathways_to_Health_Disparities

disparities in health will be partially contingent upon our ability to address inequality in the current environment."

For the most part, this is not being done. Healthcare is often an "after the fact" industry, treating bodily disease as oppose to the "upstream" impacts on the body. That's not surprising. Health practitioners can only do so much. They can treat sick individuals, but sick societies? That's not up them. It's up to "us". And while the field of SDOH is increasingly answering the bell, the insights needed to make effective change are just getting started.

END