PREDICTIVE MODELS FOR SOCIAL DETERMINANTS OF HEALTH IN KP MEMBERS AND COMMUNITIES:

AN ISSUE BRIEF FROM KAISER PERMANENTE'S SOCIAL NEEDS NETWORK FOR EVALUATION AND TRANSLATION (SONNET)

Kaiser Permanente (KP) Primary Authors:

John F. Steiner MD MPH (Institute for Health Research, KPCO), John L. Adams PhD (KP Center for Effectiveness & Safety Research), Drew Clausen PhD (KP Health Innovation Studio, KPSC), Kathryn M. Clift (Utility for Care Data Analysis, KP), Alyssa Millan MPH (KP Health Innovation Studio, KPSC), Claudia Nau PhD (Department of Research and Evaluation, KPSC), Douglas Roblin PhD (Mid-Atlantic States Permanente Research Institute, KPMAS), Julie A. Schmittdiel PhD (Division of Research, KPNC), Emily B. Schroeder MD PhD (Institute for Health Research, KPCO)
INTRODUCTION

The World Health Organization defines social determinants of health as “conditions in which people are born, grow, work, live and age, and the wider set of forces shaping the conditions of daily life.” (http://www.who.int/social_determinants/en/) Social determinants are the upstream causes that put people at “risk of risks” (Glass, 2006), and that may manifest in specific basic resource needs such as unsafe housing or inadequate access to food or transportation. These non-medical needs, in turn, are critical determinants of health and the effectiveness of health care (Phelan, 2005; Braverman, 2014).

The executive leadership of Kaiser Permanente (KP) has committed to “identify and address individuals’ most pressing basic human needs as a standard part of quality healthcare and to achieve health equity,” in partnership with the communities KP serves. As part of this initiative, KP Community Health established the Social Needs Network for Evaluation and Translation (SONNET), an interdisciplinary consortium of researchers and evaluators with expertise in assessment of social needs, pragmatic interventions, and implementation science. Consistent with the KP Research Strategy, SONNET provides consultation to clinical and operational programs, promotes communication of new developments in KP, and leads projects to advance the social needs agenda of the organization. Out of the broad array of social determinants, SONNET focuses primarily on five basic resource needs (housing, food, transportation, energy/utilities, and medical costs) that underpin health behaviors and participation in health care and affect health outcomes.

In this Issue Brief, a working group of SONNET investigators and colleagues from other KP departments discusses the role of predictive analytics in identifying basic resource needs and assessing their effect on health outcomes. We define predictive analytics as the development of statistical models “combining a number of characteristics (e.g. related to the patient, the disease, or treatment) to predict a diagnostic or prognostic outcome” (Amarasingham, 2014). Throughout, we highlight approaches to conduct predictive analytics at both the individual (member) level and the community level.
PRINCIPLES OF PREDICTIVE ANALYTICS

The science of predictive analytics uses statistical tools to combine multiple characteristics ("variables") that describe a person and/or a community into a multi-variable model that describes associations with an outcome of interest. That statistical model is “predictive” when it is used to describe the status of an individual, group or community whose status is unknown. A predictive model may either be used to identify a current state (such as a current resource need), or to anticipate a state that may develop in the future.

A basic resource need may be the outcome variable for a predictive model. For example, community health leaders may want to determine whether residence in a specific zip code or census tract is associated with a higher prevalence risk of economic deprivation or financial insecurity. This outcome can be measured by neighborhood deprivation indices (Singh, 2003; Kind, 2014). Conversely, basic resource needs may also be included in predictive models as independent variables that may be associated with other clinical or operational outcomes. For example, transportation barriers may be evaluated as a predictor of missed clinic appointments in a project that evaluates whether offering a subsidized ride-sharing service might help high-risk KP members keep their appointments. (See Appendix 1).

The purpose of a predictive model for a basic resource need, much like a diagnostic test in clinical medicine, is to accurately differentiate between individuals who have that need “(true positives)” and those without the need “(true negatives”). (Figure 1) One characterization of a predictive model’s accuracy is its false positive and false negative rate. A false positive occurs when the model predicts a need in a person who does not have it. A false negative occurs when the model fails to predict the need when it is present. (Figure 1)

<table>
<thead>
<tr>
<th>Predicted condition (by statistical model)</th>
<th>True condition (by “gold standard” assessment)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total population</td>
<td>Condition positive</td>
<td>Condition negative</td>
</tr>
<tr>
<td>Predicted condition</td>
<td>True positive (TP)</td>
<td>False positive (FP)</td>
</tr>
<tr>
<td>positive</td>
<td>Predicted condition</td>
<td>False negative (FN)</td>
</tr>
<tr>
<td>negative</td>
<td>Sensitivity = TP / (TP + FN)</td>
<td>Specificity = TN / (FP + TN)</td>
</tr>
</tbody>
</table>

Although advanced statistical methods may reduce false positives, false negatives, or both, the accuracy of prediction is ultimately limited by the quality and relevance of the data used in the model. Whether a predictive model is good enough for a purpose depends on the decision to
be made or the action to be taken. To assure that a model is “fit for purpose,” model developers and decision-makers inevitably need to make trade-offs between false positive rates and false negative rates. While these tradeoffs can be based on statistical considerations, predictive models are generally more useful in decision-making when the balance between false positive and false negative findings is based on clinical or population health knowledge.

For example, a model to screen for basic resource needs in a population with an overall low level (prevalence) of social needs may need to minimize false positives so that staff members do not spend most of their time contacting members who do not have those needs. Alternatively, if the planned action is something simple, low-cost, and free of harm (e.g., targeting an email or text message campaign to a specific membership segment) a higher false positive rate may be acceptable. If the planned action is particularly expensive or risky, it may be necessary to refine the target group with additional outreach and screening questions to further concentrate the target population.

### STATISTICAL TOOLS FOR PREDICTIVE ANALYTICS

Many statistical techniques can be used to develop and validate predictive models. Statistical approaches such as logistic regression or Cox proportional hazards models are well established in clinical and epidemiological research. In these models, choices about variables to include are made by the development team. The resulting models produce “transparent” results that identify all the variables explicitly and estimate the strength and direction of their association with the outcome of interest. This transparency helps decision-makers understand how these models were produced and enhances the “face validity” of the model.

Such models demonstrate associations between variables but cannot determine whether the predictor variables are causes of a particular outcome. When the model contains predictors that are potentially modifiable, interventions to address those predictors may help establish causation. For example, if self-reported food insecurity is associated with difficulty in shopping or preparing meals, the need may be addressed by providing transportation assistance or enrolling the individual in a supplemental meal program.

Newer statistical tools for predictive analytics are collectively referred to as “artificial intelligence” or “machine learning” approaches. These approaches are widely used in the technology sector but remain relatively uncommon in health care. In principle, they do not represent a radical departure from traditional statistical models, but involve less human guidance and more machine guidance (Beam, 2018). Their appeal is based on the hope that they will be more accurate (i.e., will simultaneously reduce false positive and false negative predictions, or both), by including non-clinical
data sources to measure both community characteristics and individual behavior (e.g. purchasing patterns) in non-clinical settings. When little is known about factors that predict a social or clinical outcome, or when existing predictors are only weakly associated with that outcome, these exploratory models can identify influences that could be further tested.

Artificial intelligence and machine learning approaches are more flexible than conventional models because they do not require a pre-specified model structure or make assumptions about the underlying data structure. Instead, most machine learning approaches apply an algorithm to a large set of predictors to build a model that can account for complex relationships between variables. These algorithms are often “self-training”, in that they learn progressively over many iterations. These models do not explicitly describe the strength and direction of association with the outcome of interest, but can provide a list that ranks predictors by their relative importance in predicting the outcome. Careful processing of results and further modeling steps are necessary to understand the modifiable factors that can be targeted by interventions. If the causal relationships within the model and the remediable factors that can be targeted by interventions are unclear, clinicians and other health care decision-makers may be suspicious of their predictions, despite their potentially greater accuracy.

Advocates claim that these newer approaches to predictive analytics will revolutionize decision-making in healthcare. They suggest that healthcare organizations will be able to make more accurate predictions ranging from hospital readmission rates and chronic disease outcomes to overall utilization of healthcare services. Applied to social determinants of health, these models could be used to predict which members or patients might have a social need. Alternatively, social determinants of health could enhance the ability of healthcare organizations to stratify patients more accurately than is possible using claims, census and consumer data alone. Advanced analytics may be particularly powerful in KP because of our ability to include the rich clinical data from our electronic health records, administrative, and enrollment systems.

Although these models are largely touted for their benefits to health care delivery systems, they could also enhance planning of community health interventions. If we can confidently predict which individuals will face food insecurity and where they live, KP Community Health could more effectively target its investments in community food resource infrastructure to the most vulnerable jurisdictions.

**PREDICTIVE ANALYTICS TO IDENTIFY BASIC RESOURCE NEEDS**

Predictive analytics can be informative in most of the five broad categories, or “pathways” by which KP can address basic resource needs in its members and its communities. (Figure 2).
Figure 2: Pathways to identify basic resource needs in individuals and communities

Clinical Pathways

Pathway 1: Clinical Care
- Data: patient report
- Prevalence: 100%

Pathway 2: Screen Large Populations
- Data: surveys, EHR
- Prevalence: Low

Pathway 3: Screen High Risk Groups
- Data: surveys, EHR
- Prevalence: High

Community Pathways

Pathway 4: Hot Spotting
- Data: surveys, EHR, administrative
- Prevalence: Very High

Pathway 5: Identify Vulnerable Communities
- Data: community level data
- Prevalence: Low to High

Predictive Analytics (Clinical)

Predictive Analytics (Community)

Connect to Community Resources to Address Basic Needs

Develop Resources for Communities
Clinical pathways

A clinician may identify a member’s basic resource needs directly during a health care encounter (Pathway 1). For example, s/he may learn of a transportation barrier when a patient apologizes for being late to a clinic appointment. This “one to one” pathway does not require predictive analytics because the member identifies her/his individual needs.

KP can complement individual case-finding by using population health strategies that have enhanced our delivery of clinical preventive services and chronic disease management. Pathway 2 demonstrates one such approach, universal screening of large and unselected populations. Currently the largest universal screening program in KP is the Medicare Total Health Assessment (MTHA) (Steiner, 2018), which has been administered to over 100,000 elderly KP members in conjunction with Medicare-mandated Annual Wellness Visits. The MTHA survey asks about food insecurity and social isolation. As noted earlier, a potential problem with this pathway is that screening surveys administered to a large population with a low prevalence of needs may result in a high rate of false positive responses. Additionally, survey responders may not be representative of the individuals or communities that might benefit most from interventions to address social determinants on health. For example, patients with financial insecurity may be reticent to disclose non-medical aspects of their lives unless they know how the information will be used and who will have access to it.

Survey results may be combined with data from other KP and community sources to develop predictive models that identify high-risk subgroups. The MTHA screening program in KP Colorado (KPCO) found that the prevalence of food insecurity was 5.7% in over 50,000 elderly KPCO members. A predictive model that combined MTHA data with EHR and utilization data showed that the highest-risk quintile (20%) of members had a 14.3% risk of food insecurity, almost 3-fold greater than in the population as a whole (Steiner, 2018). Once such models have been demonstrated to have acceptable false positive and false negative rates, targeted assessment of the highest-risk group (Pathway 3) can potentially replace population-wide screening.

Pathway 3 promotes targeted screening of high-risk sub-populations, such as members with medical complexity, high costs, or financial barriers to care (such as members receiving Medicaid benefits). This has been the most common approach to identify basic resource needs in KP members. In this pathway, predictive modeling may further refine the identification of high risk subgroups. For example, predictive models could describe the subset of high-utilizing members whose frequent emergency department visits are related to hunger or social isolation, distinct from those whose visits are driven by an advanced illness.

Some KP investigators are exploring the potential of KP data systems to identify subtle manifestations of member distress. In KP Southern California, Claudia Nau PhD is studying novel use-of-care patterns, patient characteristics and behaviors that may be associated with socioeconomic
distress and that can be derived from information in the EHR and membership files. Examples include frequent missed outpatient appointments, repeated, short distance changes in home address or telephone number, low medication adherence, and residence in a neighborhood with low average income.

**Community pathways**

While interventions such as referrals to social services are useful in resolving individual needs, they do not address the “upstream” social and environmental conditions that lead to basic resource needs. Identifying geographic clusters of members with a high prevalence of social needs can help KP partner with community organizations to build community capacity and promote long-term, structural change. Identifying geographic clusters of risk can also help KP Community Health efficiently concentrate resources in specific KP medical office buildings and their surrounding communities.

In **Pathway 4**, vulnerable communities or other geographic areas can be identified using individual-level data from KP members, commonly known as “hot-spotting” (Clift, 2014). For example, geographic information systems (GIS) may identify neighborhoods with high rates of chronic respiratory disease exacerbation related to substandard housing or environmental exposures such as air pollution. This approach may not require complex statistical models, but it can identify at-risk communities to explore through **Pathway 5**.

In **Pathway 5**, publicly available, community-level data from sources such as the US Census, crime statistics, and information on access to healthy food can be used to target geographic areas where basic resource needs are more prevalent in all residents, including but not limited to KP members. Information from KP-sponsored Community Health Needs Assessments may also be helpful in this effort. Well established analytic tools can combine individual-level and community-level data to develop individual predictions. Such approaches could identify KP members who live in a “hot spot” of need but whose individual characteristics are predictive of greater or lesser need than their neighbors. For example, children with asthma who live in a community with high exposure to fine particulate matter (PM$_{2.5}$) may have more frequent exacerbations that could predict emergency department utilization or prescription of “rescue” inhalers. Some regression models can incorporate both individual and community level factors, and test for interactions between them. Other analytical methods such as exploratory factor analysis or cluster analysis may be useful in identifying which individual and community features are related.
DATA SOURCES FOR PREDICTIVE ANALYTICS IN KP

Individual data on KP members

Data on individual KP members is available within the delivery system and from public, non-medical sources. The Virtual Data Warehouse (VDW), a federated repository of comprehensive, high-quality data used by all KP research departments, has not been widely used for operational analytics, but is a data source worth exploring for operational use (Ross, 2014). Individual data about basic resource needs can be obtained from KP-developed surveys such as the MTHA or the Your Current Life Situation (YCLS) survey. In KP Northwest, needs identified through such surveys are translated into ICD-10 social codes that can be added to problem lists or attached to specific clinical encounters (Torres, 2017; Friedman, 2018). In its 2018 release, Epic, the KP EHR vendor, has added a module to record social determinants of health that is under evaluation by KP regions.

Additional data on individual members who struggle with paying their portion of health care costs could be derived from KP’s Medical-Financial Assistance (MFA) programs, or by identifying members with unpaid account balances. The use of natural language processing (NLP), which analyzes free text from clinical notes, imaging reports and other sources, is another frontier for identifying variables to include in predictive analytics within KP. Some operational groups in KP have begun to gather publicly available, individual-level data on KP members to assist in their analytic activities. Examples of these data sources include credit reports and automobile registrations.

Community level data

Many operational and research groups within KP already have access to community-level data. Some of the available information is summarized in Appendix 2. For example, the KP Utility for Care Data Analysis (UCDA) maintains current and historic geocoded home addresses for KP members. This information is further enriched with community-level data from the US Census, other federal agencies (e.g., US Department of Housing and Urban Development), and commercially-available datasets. This detailed location information can also be used to evaluate a member’s proximity to hazards and to estimate travel times to resources. All this contextual information enables useful insights regarding the challenges faced and support available to members based on their location. In addition to providing contextual data about a member’s community, the UCDA also produces “hot spot” analysis for specific disease prevalence, utilization, and preventive measures (Clift, 2014). This type of clustering analysis can help inform clinical, population, and community-based intervention strategies.
LIMITATIONS OF PREDICTIVE ANALYTICS

The potential benefits of predictive analytics in KP must be assessed in full awareness of their limitations.

1. Controversies in other industries have demonstrated the ambivalence that many consumers have about the privacy of their personal data. These concerns are amplified when the data in question relate to their personal health conditions and care choices. Before KP embraces the use of individual-level, non-clinical data in developing its predictive models, we should assure ourselves that our members, in general, support that effort. To date, the attitudes of KP members toward the use of their data in predictive analytics have not been systematically assessed. A useful test is to ask, “If our members knew that we were using this information, would they approve?” Some members may view the use of such models as discriminatory or a form of stereotyping.

2. Earlier, we drew an analogy between predictive analytics and clinical diagnostic testing (Figure 1). Like diagnostic tests, predictive analytic models can have unintended consequences (Shah, 2018). “False positive” results may trigger unnecessary assessments of members for needs that they do not have, while “false negative” results may lead KP to overlook the resource needs of members who are struggling with medical and social conditions. These unintended consequences must be considered in planning for predictive analytics. Decision-makers should also avoid over-reliance on predictions made by any statistical model when they have additional information to guide an individual decision (Smith, 2013).

3. Any statistical model provides more accurate predictions for groups than for any individual. As consumers, we recognize that companies in other industries can predict some of the goods we will enjoy, but their suggestions are not always aligned with our tastes and interests (Ellenberg, 2014). Quantitatively stated, individual-level models have wider confidence intervals than group predictions. This issue is a well-known limitation of predictive analytic tools such as the Diagnostic Cost Groups (DCG), which KP and other health care systems and health insurers have long used for actuarial purposes. DCGs can accurately risk-adjust payments for likely variations in member costs, but are inadequate to make predictions at the member level or to target interventions. This inherent limitation affects all analytic tools, including artificial intelligence or machine-learning based models. Thus, all analytic models may be more beneficial for population and community health purposes, such as projecting the number of community resource specialists necessary to address the needs of the Medicaid members in a KP region or identifying neighborhoods that lack access to healthy food, than they are for predicting whether an individual member has a specific need.
4. All statistical models are vulnerable to potential biases. For example, models based on KP EHR data incorporate substantially more information from members who use the system regularly than from those who do not. As a result, an EHR-based model will under-represent the experience of infrequent users, who may differ in important ways from regular users.

5. Even models based on large total samples may inaccurately predict the experience of small but important subgroups. This concern for members of minority groups that experience disparities in care has led to calls for “algorithmic justice” (Char, 2018, https://www.ajlunited.org/). These biases are easier to identify in “transparent” models where all variables are available for scrutiny than in “opaque” models where the variables may not be known (Gianfrancesco, 2018).

6. The basic resource needs of individuals or communities are not static. Economic fluctuations may dramatically affect individual needs, while communities may re-purpose abandoned areas into green space or recreational space over time, improving aesthetics and walkability. Thus, periodic reassessment of basic resource needs at both individual and community levels is paramount.

7. Finally, individuals are not necessarily associated with one and only one social context. Community-level data available to KP permits associations of individuals with residential contexts. But children have social needs that might be determined independently or jointly by their school environment, while employed adults have social needs that might be influenced by their worksite. Thus, concurrent measurement of multiple community contexts must also be acknowledged as challenges for predictive modelling.

In Appendix 2, we provide an example of the use of predictive analytics to address social determinants of missed outpatient appointments in primary care, a common operational concern for KP. The example is based in part on an existing predictive model, and in part on work in progress within KPCO. This work represents a hybrid of Pathway 3 and Pathway 5.

**CONCLUSIONS AND RECOMMENDATIONS**

In conclusion, this SONNET working group proposes five recommendations for KP to consider as we develop expertise in advanced predictive analytics to address the basic resource needs of our members.

1. KP should embrace advanced predictive analytics for the benefit of our members, the health care system and the community.
2. Developers of predictive analytic models should engage relevant stakeholders, including KP members and community residents, decision-makers in KP and the community, and frontline staff from KP and community organizations. These stakeholders should be included in the design process, where they can contribute to formulating the analytic question and reviewing candidate variables for predictive models. They should also participate in interpreting findings and designing interventions. The ethical dimension of predictive analytics should be in the foreground of these collaborations (Amarasingham, 2014; Amarasingham, 2016). Surveys of member attitudes with respect to data privacy and use in predictive models, the need for informed consent, and the acceptability of specific sources of data may also help KP develop analytical models that address operational needs while respecting member concerns.

3. Predictive analytic models should be rigorously evaluated once they are developed (Shah, 2018). Models developed in one population, such as a specific KP region, should be validated in other settings (Schroeder, 2017; Karter, 2017). The impact of the models on actual decisions and outcomes should also be assessed. These evaluations should estimate false positive and false negative rates, and assess the implications of both types of error for individual care and population health.

4. KP should continue to develop internal capacity in predictive analytics. We should incorporate additional data sources and continue to develop skilled analytic teams. Bringing external data into KP will reinforce our commitment to maintain the confidentiality of our members’ information, and will help us tailor the resulting decisions to the needs of our members and communities. By doing so, we may also be able to avoid legal and reputational issues that have affected other industries which are heavily invested in predictive analytics (Duhigg, 2012). Expertise to analyze these data resides in KP’s research departments and in operational groups across the organization. Investments in both settings will further enhance this capacity.

5. As KP continues to develop internal capacity, we should also consider developing relationships with external vendors that maintain large data libraries and expertise in predictive analytics. These collaborations will be advantageous in areas where internal expertise or capacity is lacking, and where external vendors maintain unique data sources. Similar to our internal learning communities, these external relationships should be collaborations among KP members, operational and clinical leaders, internal data experts, and the vendor. The approach to vendor relationships that is being taken by the Social Services Resource Locator provides a good model for such joint ventures.
Operational context

Missed appointments are common in primary care settings in KP and other health care systems. Missed appointments deprive members of needed care and are inefficient for the delivery system. Interventions to reduce missed appointments have generally involved pre-visit reminders by text message, automated interactive voice response (IVR) calls, or “live” calls from clinic staff.

Researchers and operational leaders in KPCO conducted an intervention in 2016 to assess whether two reminder calls were more effective than a single reminder in reducing missed primary care appointments. Over a two-month period, 54,066 KPCO members were randomly assigned to receive a single text or IVR reminder one day prior to their visit (the existing approach), a single reminder delivered three days prior to the visit (which might allow the clinic more time to schedule a new patient into a vacated appointment), or two reminders three days and one day prior to the visit. The trial found that 5.8% of members missed their appointment after a three-day reminder, 5.3% missed after a one-day reminder, and 4.4% missed after receiving both reminders (p<0.001) (Steiner, 2018).

Development of predictive model

Using data from the KPCO appointment system and electronic health record, researchers developed a transparent, logistic regression-based predictive model for missed primary care appointments in these 54,066 members. This approach reflects Pathway 2 in Figure 1 of this Issue Brief, since the model was developed on an unselected population of members to identify the subgroup at highest risk.

Unsurprisingly, the strongest predictor of a missed visit was the number of prior missed visits. Several socioeconomic variables were also significant predictors: younger age, male sex, black or unknown race/ethnicity, single or unknown marital status, and Medicaid insurance. The model was highly accurate (c-statistic = 0.93), and strongly differentiated members in the upper 25% of risk, who missed 23.3% of their appointments, from those in the lower 75% of risk, who missed 0.4% of their visits.

Planned predictive analytic activities

Given this unexpectedly high rate of missed appointments in a segment of KPCO members, operational leaders and researchers in KPCO are now planning to use predictive analytics to identify remediable barriers to appointment-keeping in this high-risk group (Pathway 3). The strategy for this work is shown in Figure 3.
We are particularly interested in assessing whether transportation or medical care cost barriers contribute to the high rate of missed appointments, since these barriers can be addressed with specific interventions. To identify transportation barriers, we will assess geographic clustering of missed appointments by clinic site or member residence, and community-level transportation data such as accessibility of public transportation (Pathway 4). To identify cost barriers, we will develop a predictive model that includes health care coverage through Medicaid, a subsidized exchange plan, receipt of medical-financial assistance, or an unpaid account balance. If these variables are associated with even higher rates of missed appointments, referral to KPCO financial counselors may be of value.

Predictive models can often benefit from confirmation using other data. In this project, we will consider administering the Your Current Life Situation (YCLS) survey, a brief, KP-developed survey to assess transportation and other basic resource needs, to a sample of high-risk members who miss appointments and comparison groups of members who did not miss their visits and those who missed visits but were classified as low-risk by our first predictive model. If these analyses identify greater transportation barriers in high-risk members, it would further confirm the potential value of offering ride-sharing services to assist these members in keeping their appointments (Chaiyachati, 2018; Chaiyachati, 2018). The ultimate value of a ride-sharing intervention would best be tested with a randomized trial or other strong evaluation design.
### APPENDIX 2.

**REPRESENTATIVE COMMUNITY-LEVEL VARIABLES IN CURRENT USE FOR PREDICTIVE ANALYTICS IN KP**

<table>
<thead>
<tr>
<th>Category</th>
<th>Social needs markers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Neighborhood socio-economic characteristics</strong> <em>(Source: Census Bureau)</em></td>
<td>% High school graduates in neighborhood (census tract)</td>
</tr>
<tr>
<td></td>
<td>% Population with graduate degree</td>
</tr>
<tr>
<td></td>
<td>% White population</td>
</tr>
<tr>
<td></td>
<td>% Hispanic population</td>
</tr>
<tr>
<td></td>
<td>% African-American population</td>
</tr>
<tr>
<td></td>
<td>% American Indian population</td>
</tr>
<tr>
<td></td>
<td>% Asian population</td>
</tr>
<tr>
<td></td>
<td>% Pacific Islander population</td>
</tr>
<tr>
<td></td>
<td>% Other Races, population</td>
</tr>
<tr>
<td></td>
<td>% Multiple Races population</td>
</tr>
<tr>
<td></td>
<td>Median household income</td>
</tr>
<tr>
<td></td>
<td>Median value of homes</td>
</tr>
<tr>
<td></td>
<td>% population enrolled in SNAP</td>
</tr>
<tr>
<td></td>
<td>% population on public assistance</td>
</tr>
<tr>
<td></td>
<td>% population on Medicaid</td>
</tr>
<tr>
<td></td>
<td>% population (pop) living below 50% of the poverty line</td>
</tr>
<tr>
<td></td>
<td>% population living below 200% of the poverty line</td>
</tr>
<tr>
<td></td>
<td>% population owning their house</td>
</tr>
<tr>
<td></td>
<td>% population living in housing with more than 1 person per room</td>
</tr>
<tr>
<td></td>
<td>% of houses vacant</td>
</tr>
<tr>
<td></td>
<td>% male population in managerial professions</td>
</tr>
<tr>
<td></td>
<td>% female population in managerial professions</td>
</tr>
<tr>
<td></td>
<td>% households with children led by single mothers</td>
</tr>
<tr>
<td></td>
<td>% population divorced</td>
</tr>
<tr>
<td></td>
<td>% population older than 65 years</td>
</tr>
<tr>
<td></td>
<td>% population moved in past year</td>
</tr>
<tr>
<td></td>
<td>% population born in US</td>
</tr>
<tr>
<td></td>
<td>% population that speaks English only</td>
</tr>
<tr>
<td></td>
<td>% population commuting by car</td>
</tr>
<tr>
<td></td>
<td>% population commuting with public transportation</td>
</tr>
<tr>
<td></td>
<td>% population walking to work</td>
</tr>
<tr>
<td></td>
<td>% population biking to work</td>
</tr>
<tr>
<td></td>
<td>% households not owning a car</td>
</tr>
</tbody>
</table>
| **Neighborhood food access**  
(Sources: Esri Business Analyst, USDA Food Access Atlas) | Food desert, urban areas: % population with no grocery store within 1 mile  
Food desert, rural areas: % population with no grocery stores within 10 miles  
Number of grocery stores in neighborhood (census tract)  
Number of convenience stores  
Number of chain fast food restaurants  
Number of pizza places  
Number of coffee shops  
Walkscore ©  
Neighborhood level spending on food categories (e.g. fast food, fruits and vegetables, soda etc.)  
Neighborhood level Personal and Property Crime Index (CrimeRisk© Index) |
| --- | --- |
| **Transportation (Source: US EPA)** | Transit service frequency  
Accessibility index - transit to jobs  
Accessibility index – auto to jobs |
| **Socioeconomic environment (Source: US EPA)** | Housing units per acre  
People per acre  
Jobs per acre  
Employment |
| **Environmental exposures (Source: California EPA)** | Ozone concentration  
Diesel emissions  
PM 2.5 (small particulate exposure)  
Toxic releases  
Traffic density  
Hazardous waste  
Solid waste  
Pollution burden  
Drinking water contaminants  
Groundwater threats |
| **Environmental exposures (Source: NOAA)** | Drought intensity |
| **Crime (Source: Esri)** | Personal crime rate  
Property crime rate |
| **Government spending (Source: OMB)** | Government spending by zip code in multiple categories |
REFERENCES


Duhigg C. How companies learn your secrets. https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Ellenberg J. What’s even creepier than Target guessing that you’re pregnant? http://www.slate.com/blogs/how_not_to_be_wrong/2014/06/09/big_data_what_s_even_cREEpier_than_target_guessing_that_you_re_re_pregnant.html


