**Multilabel and Multiclass Analysis of Social Determinants of Health**

**Francisco Iacobelli and Rufino Salgado**

Northeastern Illinois University

5500 N. St. Louis,

Chicago, IL, 60625

**Abstract**

For many people, challenges in receiving treatment for illnesses begins before they ever set foot in a doctor’s office. Many of the socioeconomic factors or lifestyles choices individuals face daily has already impacted their access to healthcare services. The goal of this paper is to evaluate patient demographic features and their relationship to barriers patients are most likely to face in accessing healthcare as well as actions providers are most likely to recommend to reduce the impact of said barriers. Our study’s results suggest a significant correlational relationship between primary languages and these barriers and actions. However, improvements to dataset consistency as well as an increase in the number of data points would be the next goals to improve upon the precision of the predictive classification model.

**Introduction**

In today’s age of mass data collection, the most private and personal of this data originates from our interactions with the healthcare system. The highly descriptive nature of this data, including demographic information and health information, makes it particularly suitable for use in predictive machine learning focusing on patients. There have been several efforts to utilize this health data in predictive machine learning modeling, primarily focusing on classifying disease from symptom descriptions, imaging data, and testing results. (Xiaoxuan Liu et al., 2019). These studies have the goal of improving diagnostics or treatment plans once individuals become patients. However, many factors aside from medical treatments contribute to the successful recovery of a patient before they ever set foot in a doctor’s office.

These barriers to healthcare access can come from any, if not several, sources in a person’s life. For some, fear or distrust of the healthcare system stemming from negative experiences or cultural attitudes involving doctors or hospitals can convince them to avoid seeking treatment in the first place. Those who do not share these fears could still find commuting to a healthcare facility too burdensome or costly, limiting their access to healthcare. If there is a language barrier, this might lead to difficulty in communicating symptoms or understanding treatment. The administrative processes involved with healthcare can impede the ability for individuals to find or schedule a doctor’s visit, especially when they need to simultaneously navigate other bureaucratic systems to do so.

These barriers are linked to the socioeconomic status and cultural attitudes of an individual. Therefore, collecting this data from individuals could potentially be leveraged to predict what barriers they would face in the future. The goal of this paper is to evaluate the viability of this by applying the demographic features of patients towards the construction of a machine learning model, specifically a multilabel classification model. If this classification model demonstrates the capability to reliably predict the obstacles a patient might face when interacting with healthcare providers, steps could be taken preemptively to lessen the negative impacts of these limitations. Foresight regarding potential barriers could enlighten healthcare providers as to what actions, or measures, would be the most beneficial to invest in to best support their patients.

**Background**

As stated previously, there has been a growing interest in the medical community on the promise of machine learning and its potential contributions to the medical field. A systematic review of studies suggests classification models meant to detect or diagnose disease these programs can perform on par with healthcare professionals’ opinions. (Xiaoxuan Liu et al., 2019) However, limitations on data collection and in external validation are factors to consider when comparing accuracies between professionals and predictive modeling. The data used in evaluating these models primarily involved specific health information of patients collected previously and focused on detecting specific illnesses.

In other studies, machine learning models were constructed using more general socioeconomic information about patients with the goal of predicting a variety of health outcomes. One study, completed by Chen et al (2020), sought to estimate probabilities of hospitalizations based on social determinants of health such as air quality and education. The authors of this study concluded that their results demonstrated the feasibility of using non-health data to predict hospitalizations. Their successes demonstrate how relevant and influential these factors are to individuals’ health.

Social determinants of health are already being looked at as a focal point for determining disparities in healthcare access. Articles reflecting on these disparities point out the impact that living conditions or economic inequality are some of several features of a group to consider as associated with negative healthcare experiences or treatment. (Penman-Aguilar et al, 2016; Koh et al, 2010) Because of the role these social determinants can play in the health outcomes of entire populations of people, many efforts by health organizations have attempted to address inequalities in healthcare access for populations according to social determinants, with the hopes of improving healthcare for all who are part of these groups.

Studies have shown evidence of the improvements in health for groups possible through interventions being set into effect, guided by these social determinants of health. (Williams DR, 2008) Improvements to living conditions or policy changes to promote home visits by medical professionals have resulted in positive changes and effectively reduced the disparities in health conditions present in populations prior to these interventions taking affect.

The impact an individual’s socioeconomic and demographic traits play such a direct role in their health outcomes, health organizations have tried to develop models of determinants for diseases which include these traits. The University of Chicago’s Center for Interdisciplinary Health Disparities Research (CIHDR) developed a model of health determinants which contribute to breast cancer diagnosis including societal factors such as race and poverty among more biological factors like genetics and mental health. (Gehlert et al, 2008) More importantly, their model does not just include social determinants, but emphasizes their interrelatedness when determining their overall effect on an individual’s health outcomes.

Our goal is to use social determinants of health such as these to predict a patient’s potential experiences with healthcare services, focusing on forecasting barriers to healthcare access and the interventions needed to manage them. The hope is that developing a method of predicting these barriers will allow providers to promote better healthcare experiences, reduce health disparities and, ultimately, create better health outcomes. These predictions will come in the form of multilabel classifications produced by machine learning models training on collected patient socioeconomic and demographic information.

The importance of taking various social determinant factors into account when making these predictions is demonstrated by the models such as CIHDR’s downstream model of health determinants. Efforts to reduce inequality in healthcare access due to widespread inequalities in society or governing bodies is important work. However, not all people affected by inequalities in their socioeconomic status will see the benefits of these efforts. Also, not every individual is affected to the same degree by every health determinant they are impacted by. Finally, as the CIHDR’s model suggests, the combination of several of these determinants could play a larger role in a person’s health than any single one. Building a predictive model that considers several of these factors in conjunction for an individual patient and predicts the most impactful barriers to their healthcare could provide a powerful tool for healthcare providers. It would allow them to address a patient’s unique set of social determinants and therefore recognize what most impactful measures they could take to reduce health disparities for that individual.

**Methodology**

The goal of this project was to take sets of patient demographic information from two sources (DuPage County healthcare providers and the Chinatown neighborhood healthcare providers). DuPage and Chinatown both provided patient data as two sets. One included a patient’s socio-demographic information such as their age, income range, and education level among other. The other set tracked a patient’s history with barriers to access and the actions taken in response. DuPage provided preexisting demographic data from 478 unique patients and tracking data from 435 patients. Chinatown provided more recent data from patients totaling to 275 unique patients’ demographic information and tracking records for 331 patients.

The first task was to consolidate tracking and demographic information into a single file with all data for each unique patient listed together. Among the tracking information for DuPage county, the data input seemed to limit one pair of a barrier and an action at a time. As a result, data entry required entering the same patient multiple times, with multiple dates, to record multiple barriers and actions they have encountered. Because of this, the tracking information had to be reformatted so that individual patients had all barriers they faced over all visits listed in one new field. The same was done for interventions done for these patients. Datasets were merged into one complete DuPage county dataset with each data point referring to a unique patient. Data was excluded in cases where a patient was not included in both datasets. In total, this produced complete data for 400 unique patients.

Chinatown recorded medical history, including barriers and interventions, already listed together for individual patients (tracked by a unique Record ID), so there was no need for reformatting in that regard. However, barrier or actions taken were separated by visit rather than listed together, such as it was in DuPage data at this point. Each patient had ten visits as a max recorded with barriers and interventions recorded during that visit. To end up with a similar list as DuPage of all barriers and interventions for a patient, the data was looped through checking each visit’s date of service. Only if a date were listed would the corresponding barrier and interventions be added to the cumulative list. These two lists were stored under newly created fields. As before, patients which were completely missing from one of the datasets were excluded from the merged dataset. This resulted in a total of 274 unique patient data points.

The next task was to combine these complete datasets into one including all patients from both regions. The challenge with this was that both regions did not collect the same categories of data or, if they did, data was formatted differently. For instance, barriers to healthcare in DuPage data were input as a letter and number code (e.g., LI1, F1, etc.) while Chinatown listed then as spelled out terms (e.g., Listener/Interpreter, Fear, etc.). DuPage data included a field for Primary Language while Chinatown data elaborated on language data by also including a patient’s fluency with English and an option to include specify what form of Chinese they spoke. For the most part, patient attributes in DuPage data appeared as PD and an abbreviation of the attribute it referred to (i.e., ‘Marital Status’ appears as ‘PDMSTAT’). Chinatown labelled attributes in more direct terminology such as ‘occupation’ or ‘household\_income’.

To merge these datasets, the first step was to decide on a labelling system for both datasets to have in common. This is because when the datasets were to be merged, the attribute names would have to match. Attributes in Chinatown’s dataset were converted to match equivalent fields in DuPage’s data because the ‘PD’ prefix and more concise nomenclature made them easier to manage as the project went one. If no equivalent field labels could be found, they were left as is to be discarded during merging.

Barrier and intervention codes required reaching out to the people of DuPage who provided the data. The data had all barriers and interventions formatted into codes however, a comprehensive list of the codes and their description was not provided initially making it impossible to convert the values in the Chinatown data. They were informed of the request and the additional material was provided containing this comprehensive code description list.

Next, the data values themselves needed to be reformatted to match as well. DuPage datasets was given formatted into codes when appropriate which could be referenced in a codebook they provided which provided more detailed information. As stated previously, Chinatown data started off without this sort of formatting. For instance, under the birth country attribute, the US could be input in any number of ways from ‘US’ to ‘United States’ or ‘usa’ which would all be considered different values. Therefore, for any attributes which both sets now had in common, their values were cycled through looking for equivalent values in the DuPage codebook and reformatted.

Something to consider was that while there were some direct matches for values could be found in the DuPage codebook for conversion (i.e., ‘China’ coded as ‘CN’), this was not always the case. Income levels in DuPage data were broken down into range of an additional $10,000 per level while Chinatown data varied from another $5,000 or $15,000 per level. In terms of employment, DuPage had only three categories (‘Unemployed’, ’Part-Time’, or ‘Full time’). Chinatown data broke this attribute down further differentiating unemployed from retirees, homemakers, and students. In these cases where direct conversions could not be made, values from attributes were converted to the system in which most values from both datasets could be categorized in a consistent way. For instance, employment statuses were formatted to follow the Chinatown employment categories. This is because doing the opposite would have all 88 patients labeled as ‘Unemployed’, ‘Student’, ‘Retirees’ and ‘Homemaker’ would be lumped under ‘Unemployed’ by DuPage county’s categories and all ‘Working’ labels would not be able to be categorized into ‘Part Time’ or ‘Full Time’. Values which could not be represented by the chosen system were included as a ‘None’ category.

Furthermore, any values entered in a foreign language was converted to ‘None’. Any codes which were not specified in the DuPage codebook were changed to ‘Chose not to answer’ or ‘Other’ if these were available codes, otherwise they were changed to ‘None’. Numerical fields with missing values were replaced with the average value for that attribute.

Six new fields were established for each patient: three for the top three most frequent barriers patients faced that individual patient faced and three for the top three interventions they received. It was clear when reviewing the data that most patients noted two or more obstacles when trying to receive healthcare. As described earlier, it is also important to consider these barriers in context of other existing barriers. Selecting the most frequent barriers and interventions for a patient also represents the most impactful or persistent that they faced. Therefore, the model that this data would be fed into would be built to predict multiple barriers and multiple interventions per patient. Frequencies of their appearance were counted from the consolidated lists produced earlier and the three most common values were entered into the three new barrier fields or intervention fields, respectively.

Originally, merging these two preprocessed datasets resulted in a merged dataset containing 674 unique patients and their ages, country of birth, primary language, and zip code as potential features and the six new lists per patients listing the three most frequent barriers and three most frequent interventions they encountered. Feeding this data into a multilabel classification model returned poor prediction results regardless of the combination of layers involved. It was suspected that the low number of features contributed greatly to this. This was shared with the healthcare groups along with a request for additional demographic information for patients and additional codebook information from DuPage to explain their coding, if possible. Several attributes were already in common between the DuPage and Chinatown datasets however their values could not be converted into a common system due to the DuPage data coding not being explained in the initial material they provided.

While waiting for additional, efforts were made to extract additional features from the existing data. Given the tracking history and counting dates of service, a field was calculated for number of visits associated with each patient. Patients’ specific addresses were not available, but by using their zip codes, the Google Maps API was utilized to determine the nearest hospitals to the given zip code, the distance in kilometers, the potential driving time, and the potential travel time on public transportation. The last bit of information was not always available through the API’s own limitations, but any missing values were replaced with the average transit time value.

DuPage county responded with a new version of the previous patient demographic information including new attributes. They also provided an expanded codebook explaining the value codes for the new attributes and previous ones which were missing explanations. As a result, the new merged dataset consisted of 674 unique patients and thirteen features per patient (age, primary language, birth country, marital status, education level, household size, income level, employment status, zip code, distance to nearest hospital, driving time to hospital, public transit travel time to hospital, and visit counts). There were also the six additional new fields for most frequent barriers and interventions.

It was also mentioned, with news that barriers and interventions were being counted to determine the most common, that being able to predict a patient’s visit intensity, or the range of visits they would be expected to have. Tracking information included dates of service which indicated unique encounters with healthcare services which were counted to determine the number of visits for each patient. Then, the mean of the visit counts was determined (16.94) and the standard deviation (13.06). A new field, Visit Range, was created and patients received a value of L, M, H, or VH for low, medium, high, or very high respectively. The upper thresholds for these categories were based on the standard deviation with visit counts below 3 labelled as L, between 3 and 16 as M, between 16 and 29 as H, and above 29 as VH. This brought the total number of features for use in predicting barriers and interventions up to 10 and created a new field, visit range, to predict.

**Modeling**

The data at this stage was a mix of numerical data and categorical data which needed to be transformed into a numerical form. Categorical features such as birth countries consisted of string values which could not be processed through a machine learning model or zip codes which were numerical but held to comparable meaning to one another. Each individual value was encoded into a binary feature of their own using the Pandas module get\_dummies() method, creating a new dataframe object with every value as a new feature with a one value for indices where it was an attribute of the patient and zero where it was not. This resulted in the thirteen original features expanding into 117 features.

Values which were already numerical like age and household size required feature scaling. This is because their values were typically greater than one and would therefore have a greater weight among the newly created binary features which maxed out at one. These features were scaled using the StandardScaler method from the scikit-learn module in python which standardized the values to a mean of zero, bringing most values closer to single digits. Standardization was done this way because the numerical features held different ranges of values meaning they there might a large contrast to values from different features. For instance, household members would rarely break single digits but visit counts regularly did so. Trying to make these values proportional through other methods like Min Max Scaling, which compresses values to between one and zero, would reduce smaller values closer to zero than through Standard Scaling

The barrier classes and action classes in the merged datasets were listed in three fields, one for each of the three most common values. These values were codes based on the DuPage county codebook and therefore had to be transformed like the categorical data from the features. The aim was to end up with each possible class appearing as a binary field with the three most common for each patient having ones values for their top three barriers and zeros in every other class.

The first step was creating a new list of every unique barrier which appeared in any patient’s top three. Following this, the MultiLabelBinarizer() method was used from the sklearn’s preprocessing module. This was necessary to avoid duplicate appearances of classes. The three most frequent values were matched up to the newly expanded classes and their corresponding classes were marked with ones to confirm their inclusion. This method produced 52 unique barriers and 33 unique actions patient could have had.

The classification model needed to predict three potential classes per patient. This required building a multilabel classification model which was done using the Keras API. The model consisted of nine layers. The input layer was a one-dimensional convolutional layer with a kernel width of three, ten filters, and a ReLU, Rectified Linear Unit, activation function. The ReLU function is a commonly used activation function with convolutional neural networks due to its flexibility in approximating functions with less expensive operations than other activation functions. The input shape was the number of features per patient.

The filter and kernel values were settled on following testing with increasing the values but eventually determining that greater values produced no significant benefit to the model’s performance and greatly reduced its efficiency. A one-dimensional MaxPooling layer was included after it with a pool size and stride value of two. Again, these values were the result of trial-and-error incrementing and decrementing the values to a point where they achieved the wanted effect without reducing accuracy significantly. Together with the following MaxPooling layer, the convolutional layer aided in reducing the impact of smaller values in the features and give them less weight in the final predictions. Experiments without these layers produced models with similar accuracy, but most classes individually performed poorer. Following the inclusion of the convolution and MaxPooling layers, overall performance remained similar however, accuracy increased for classes with the greatest representation.

These layers required three-dimensional array inputs incompatible with the following layers, so a Flatten layer was included to reduce the dimensionality. The remaining layers are a series of three pairs of Dropout layers and Dense layers. Each Dropout layer drops a 25 percent proportion of the incoming values. Each Dense layer has a decreasing number of nodes equal to a multiple of the number of classes. The first Dense layer has triple that amount as output nodes, the next has twice the amount as output nodes, and the final Dense layer having an equal number of nodes to classes. The output layer uses a sigmoid activation function to ensure that the value of each output node is considered independent of one another, necessary for multilabel classification. The model utilizes the binary cross-entropy loss function for determining the final class probabilities independently. In combination with the previous sigmoid activation function, this model produces a probability of each class being included in the given set of classes for a patient. It uses the Adam optimizer function.

The evaluation of the model was done through a Repeated K-Fold cross validation. K-fold cross validation allowed the repeated iterations of the data with different splits of data each time. The aim of this is to provide variety in the data trained on the model and tested by it, evaluating how well generalized the model is to future data. Unfortunately, the classes in this case were largely imbalanced. The ‘None’ barrier class, the class representing visits with no barrier encountered, represented an overwhelming number of barriers listed by patients at over 2000 while some classes appeared only a few times. Running a Repeated K-Fold cross validation at an 80-20 split of data, a common convention, would likely lack much information on many of the lesser represented classes in the results. An alternative to this method is the Stratified K-Fold Cross Validation method which generates weighted divisions of data based on the proportion of each class among all classes. However, this method is not is not able for use with multilabel classification. To use as representative of a training and testing sample as possible, the splits were divided into 50-50 sets. The repetitions were done 5 times, as the Repeated K-Fold function shuffles data, to increase the likelihood of classes being included in testing sets.

**Visit Intensity**

Prediction of visit intensity involved training the same patient features (excluding the visit count feature) on a multiclass classification model. As stated previously, classes consisted of ranges L, M, H, and VH (low, medium, high, and very high) created from the standard deviation of the visit count information. Visit count was removed from the selected features to avoid feature leakage because the visit ranges because of this. The numerical features were standardized, and categorical features were encoded into binary formats. In contrast to multilabel classification, each patient had only one class to predict each rather than multiple. Because of this, the classes of the training and testing sets did not need to be encoded as well. The data was split into training and testing sets based on an 80-20 split.

Due to the multiclass nature of the classification, Support Vector Machine (SVM) algorithms were decided on for their multiple options in dealing with nonlinear data. Three forms of SVM algorithms were tested with the data to determine the best fit for the prediction. The three included the LinearSVC algorithm, the Support Vector Classifer (SVC) algorithm with a polynomial function kernel, and an SVC algorithm with a Radial Basis function kernel.

The LinearSVC algorithm was set to a regularization value of 10 and was set to run for a max number of iterations of 1000. The SVC algorithm with a polynomial function was also set with a regularization value of 5 and a degree value of 3 to increase the flexibility of the algorithm to approximate a nonlinear function. The RBF kernel SVC algorithm was set to a regularization value of one and gamma value was 2, again, to increase the flexibility of the algorithm. The parameters where settled on following continual testing to maximize their accuracy. Their performance was compared through a merged confusion matrix displaying the numbers of correctly classified visit ranges by each algorithm. LinearSVC ultimately was chosen as the optimal algorithm based on the it being able to reliably predict instances of each class while the other two, polynomial and RBF function kernels, failed to capture instances of Very High class patients, shown in Appendix A.

**Analysis**

Analysis of the most frequently listed barriers revealed the relationships between barriers and patient features. These relationships, in the form of correlations, tell us which features played the greatest roles in causing these barriers patients encountered. Some of the most frequent barrier classes to appear included case management issues, being uncomfortable communicating in English, and encountering no barrier at all during visits. According to correlations between features and specific barriers, patients reporting being uncomfortable with speaking English as a challenge were, understandably more likely to speak primary languages other than English. However, we also see language being a strong indicator of case management challenges, with patients who listed English as a primary language being the most likely to list this barrier. Because this class is used in cases where a patient’s issues did not fit all the other possible classes. This suggests that the comfort a patient has with English can be an effective bit of information to rule out most barriers as potential issues. This also stresses the importance of providing services for patients with English as their second language.

Another strong relationship based on languages were cases where the healthcare staff lacked someone who spoke the patient’s primary language. Patients who reported this were more likely to have Spanish as a primary language other than English and Other primary languages, most likely Chinese from patients in Chinatown. Those least likely to experience this barrier were listed as Other primary language, again including Chinese speakers. These patients mostly originate from Chinatown where there might be more services prepared for Chinese speakers than they would typically find outside their neighborhood. However, this might be because Other primary language speakers regularly reported another similar language barrier, discomfort with English. These barriers seem to be describing similar obstacles faced by non-English speakers in general just divided among separate classes.

This influence by primary language goes hand in hand with many patients’ countries of birth. Naturally, Spanish speakers largely report being from Mexico and many patients with Other primary language were most likely Chinese speakers from China. One of the stronger relationships related was related to primary language and birth countries. A lack of family support was a barrier strongly related to China-born patients and the least related by Spanish speakers, possibly due to cultural differences regarding family networks or the number of generations their family may have in the US following immigration. These demographic features being so strongly related to the most common barriers indicate that the greatest potential impact to reducing barriers faced by patients would come from providing or improving services for non-English speakers.

The instances of encounters with no barriers listed was the largest of all classes yet had no strong relationship to any one feature. It indicates this class was used by a wide variety of patients and therefore could not be associated to any particular type of patient. Considering this, the fact that it referred to no challenges for patients, and how imbalanced its presence was in the data, future analysis on barriers could potentially ignore this class to focus on actual barriers encountered.

We see a similar trend of language when observing what services patients received. Looking at the correlations of the most common actions taken against the demographic features of patients, Appendix B correlations show again that the strongest determinant of the services they required was their primary language. One of the most required services was providing someone to accompany the patient to health services, and we see that those that needed it the most were Other primary language speakers. This could be a result of the degree of discomfort they reported with English making it difficult for them to navigate to and through health visits. To make matters worse, unlike Spanish speakers, they also reported lacking family support networks which could be another factor for why they needed to have someone to assist them. Spanish speakers, on the other hand, which were found to be the least likely to report a lack family support, were also the least likely to require this service, most likely because they had family ready to fill those roles.

The relationships between patients with the Other primary language trait demonstrates the effect of providing the accompaniment service for them. Looking at the correlations between actions taken and patient features in Appendix B, we see that while these patients were the most likely to be accompanied to health services, they were also the least likely to receive or require many of the other actions taken. Out of twenty-nine total recorded actions taken, patients in the Other primary language group appeared as the feature with the weakest correlation to fourteen of those. Out of 46 barriers reported, they also appeared as the least likely feature for 15 barriers.

A possible explanation for this is that providing accompaniment to health services acts as a mitigating factor for several barriers at once. Those who accompany patients can calm fears they might have navigating health services alone, they could provide transportation, help patients comprehend medical procedures, and possibly act as interpreters. This is supported by the fact that some of those barriers with the weakest connection to the Other primary language group are barriers including stress, fear of testing, misunderstanding medical terms, lack of transportation, and staff not speaking their primary language. This effect seems to extend to actions taken as well. As stated, the Other primary group, which includes Chinese speakers, had the most negatively correlated relationships with almost half of all actions taken by healthcare providers. Because they were the most likely to be provided accompaniment to health services, it is possible this also reduced their need for other actions to be taken.

Correlations between features and visit intensity shows the ratio of number of visits and languages. Medium visit intensity patients are the largest portion of patients and we see in correlations between visit ranges and features in Appendix A that if they have a Medium visit intensity, they are most likely in the Other primary language group, including Chinese. When interpreting results, it is important then to consider that the model predictions and correlations are largely based on those patients, possibly patients from the Chinatown neighborhood. We also see the patients most likely to need many visits are Spanish speakers and Mexican born. Some explanation from this may lie in the relationships with income levels represented in Appendix C. Although the correlation is not as strong as others, income levels between $10,000 and $20,000 and income levels between $20,000 and $30,000 had their strongest positive correlations with Spanish speakers and Mexican born patients. These income levels are in the bottom half of income levels recorded suggesting Spanish speaking patients and Mexican native patients analyzed have lower income levels which could be making them more susceptible to health issues from unhealthy lifestyles related to it. Exacerbating the problem for Spanish speakers is that the numbers of household members recorded as a feature also list them as their most positively correlated trait.

Looking at the results from the multilabel classifications for barriers and actions in the results in Appendix D, we see the mix of correct and incorrect predictions for each class. Certain classes, the more common barriers like None, Case management, and Uncomfortable communicating in English appear consistently in other classes’ predictions. It demonstrates how often other classes were misclassified as them. The degree of imbalance these classes led to other classes regularly being mistaken for them. The confusion might stem from the strong relationships with language stated previously. The most common barriers and actions tended to have high correlations with certain language groups and so the model may have learned to use language values as the deciding factor when detecting these classes. However, this also means those languages would play a strong indicator of the most frequent classes when appearing in a patient’s features even when the true classes are completely different.

Similarly, looking at the results from the predicted actions taken in the graph in Appendix D, we see relationships with language have the opposite effect. According to the correlations between actions taken and features revealed that those in the Other primary language group were more likely to require accompaniment to health services and less likely to receive many other actions. We see accompaniment to health services is mostly predicted correctly and, unlike with barrier predictions, other classes do not appear to be misclassified as it often. In conjunction with the mix of classes in the barrier predictions, we see languages play a large role in correctly, and incorrectly, predicting classes.

**Future Work**

For efforts to improve healthcare access based on the results from this project, language groups have shown to be a major factor in whether a patient will encounter hurdles to care. It is the greatest indicator of a patient encountering the most common barriers and of them requiring the most common actions taken in response. A language barrier has been shown to produce multiple barriers and responses to overcome this alleviate these issues accordingly. More efforts should be made provide patients with information in their own languages if not prioritizing involving staff with shared languages as the population they are working with.

As shown with patients in the Other primary language group, ensuring there was someone who could accompany patients facing these barriers effectively reduces the effect of multiple barriers with one action. At the same time, it also decreases the need for several other actions as it could potentially serve the same purpose as multiple actions. This appears to be currently done for primarily Chinatown patients but doing so for other at-risk groups in other locations may achieve the same effect.

Spanish speaking patients who were analyzed seem to largely be part of large, lower income households. Health services could be made with that in consideration. Possibly their income is limiting healthy options such as food or possibly its limiting them financially from reaching out for medical treatment earlier into their illnesses. The future efforts to decrease visits to health services should focus on providing access and information about services which they could take advantage of when they are under a significant financial burden.

For efforts to improve analysis and predictive models of patient socio-determinants of health, the focus should be on gathering more information on different barriers and actions. The data provided for this analysis contained a disproportionate amount of each. This imbalance meant only the most common barriers, actions, and visits counts could be determined from the data but at the cost of to very inaccurate results for predictions of less common classes. If it is an option, the focus should be on collecting instances of patients which reported these rarer classes of barriers or actions or visit counts for training the predictive models. One option of creating this data might be by consolidating similar classes. As the example with different language barrier demonstrates, having too many barriers to choose from can lead to patients being recorded with different barriers though they are facing similar issues which can be dealt with by the same sort of actions. Instead, barriers could be grouped to avoid confusion and in doing so would create more focused groups of patient data per class for improved modeling.

**References**

Penman-Aguilar A, Talih M, Huang D, Moonesinghe R, Bouye K, Beckles G. *Measurement of Health Disparities,* *Health Inequities, and Social Determinants of Health to Support the Advancement of Health Equity.* J Public Health Manag Pract. 2016;22 Suppl 1(Suppl 1): S33-S42.

Chen S, Bergman D, Miller K, Kavanagh A, Frownfelter J, Showalter, J. *Using applied machine learning to predict healthcare utilization based on socioeconomic determinants of care*,Am J Manag Care. 2020 Jan;26(1):26-31.

Gehlert S, Sohmer D, Sacks T, Mininger C, McClintock M, Olopade O. *Targeting Health Disparities: A Model Linking Upstream Determinants to Downstream Interventions.* Health Affairs 2008 27:2, 339-349

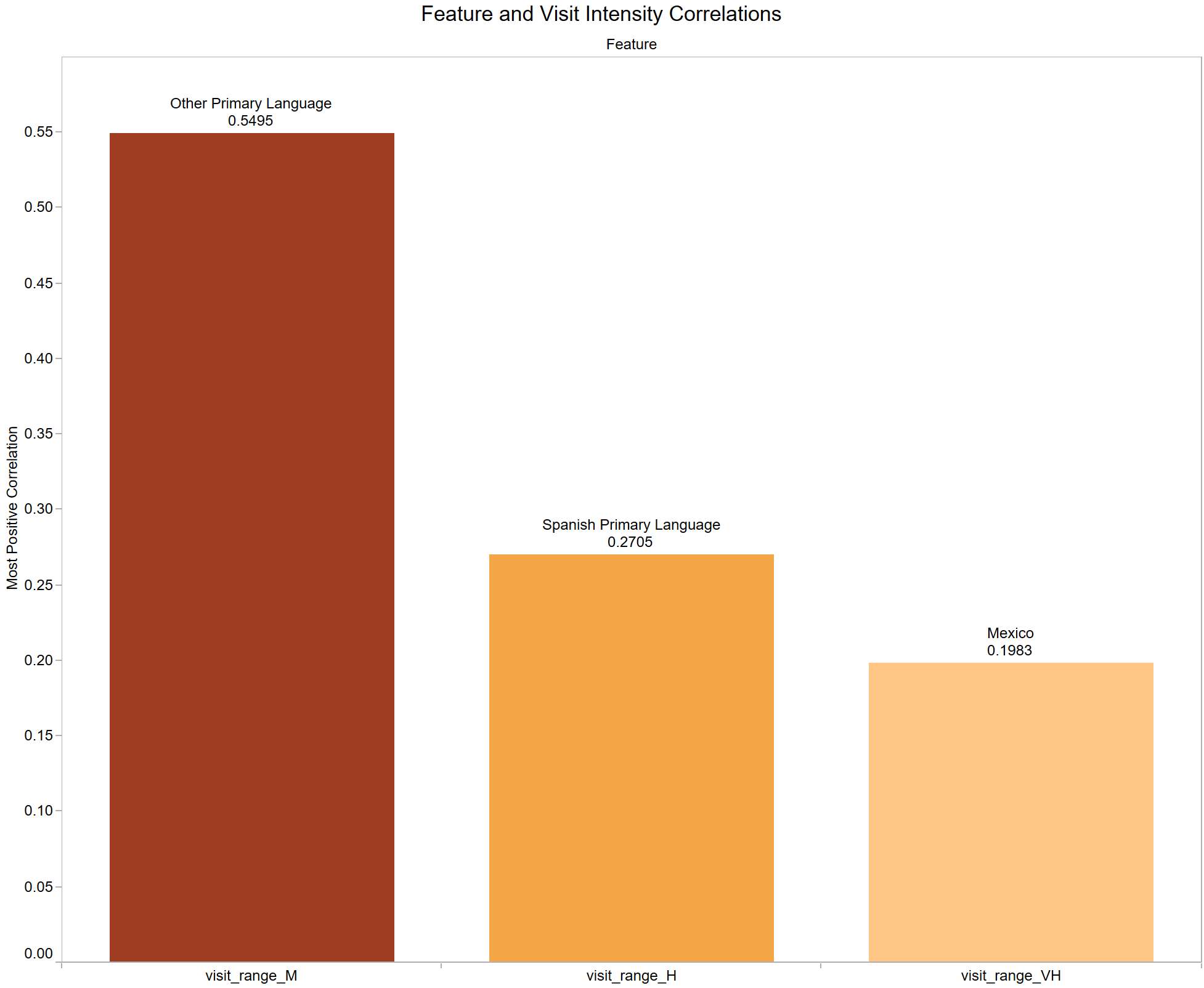
Koh HK, Oppenheimer SC, Massin-Short SB, Emmons KM, Geller AC, Viswanath K, *Translating Research Evidence Into Practice to Reduce Health Disparities: A Social Determinants Approach,* American Journal of Public Health 100, no. S1 (April 1, 2010): pp. S72-S80.

Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, Mahendiran T, Moraes G, Shamdas M, Kern C, Ledsam JR, Schmid MK, Balaskas K, Topol EJ, Bachmann LM, Keane PA, Denniston AK. *A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis*, The Lancet Digital Health, Volume 1, Issue 6, 2019, Pages e271-e297.

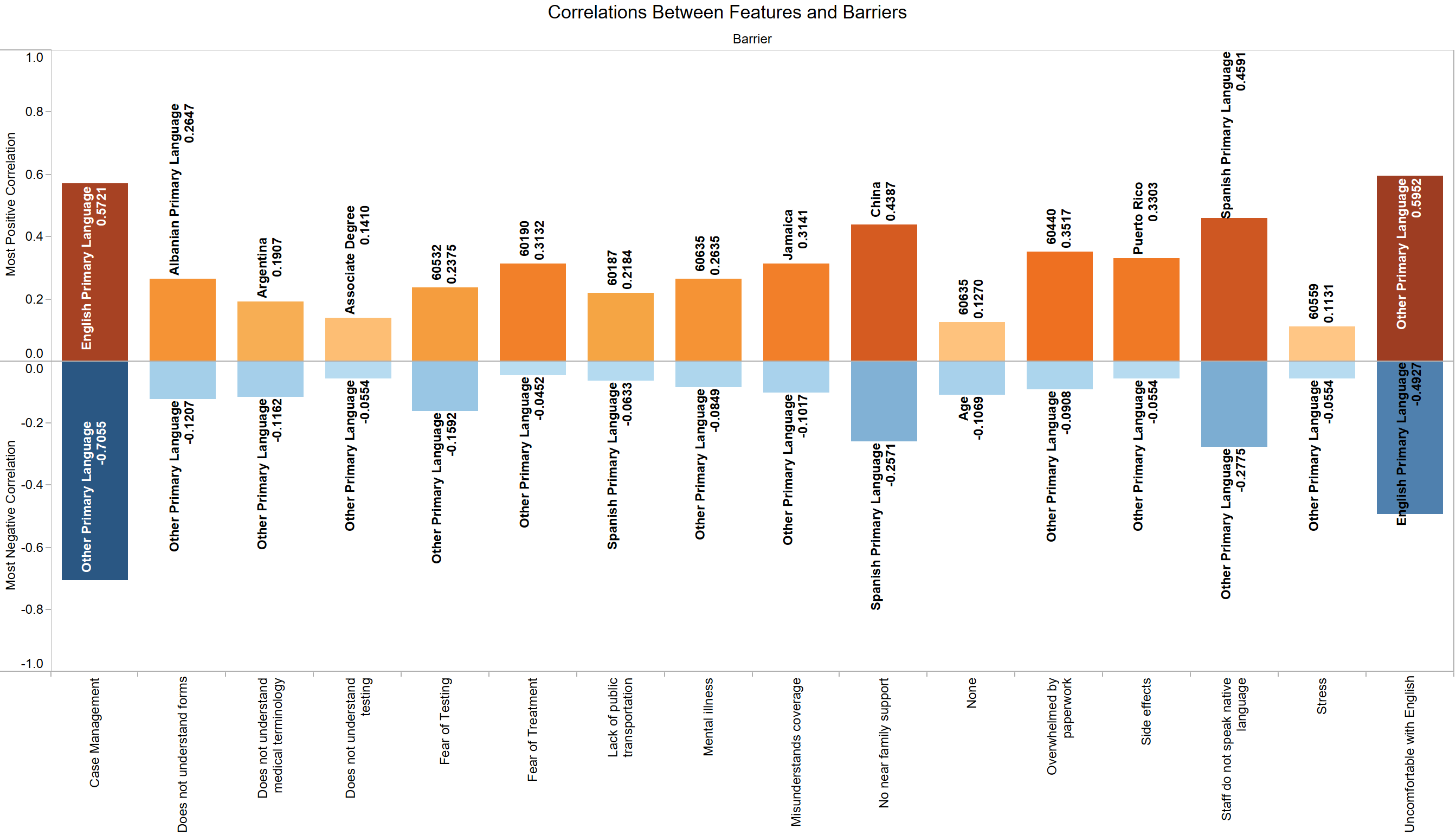
Williams DR, Costa MV, Odunlami AO, Mohammed SA. *Moving upstream: how interventions that address the social determinants of health can improve health and reduce disparities.* J Public Health Manag Pract. 2008; 14 Suppl (Suppl): S8-S

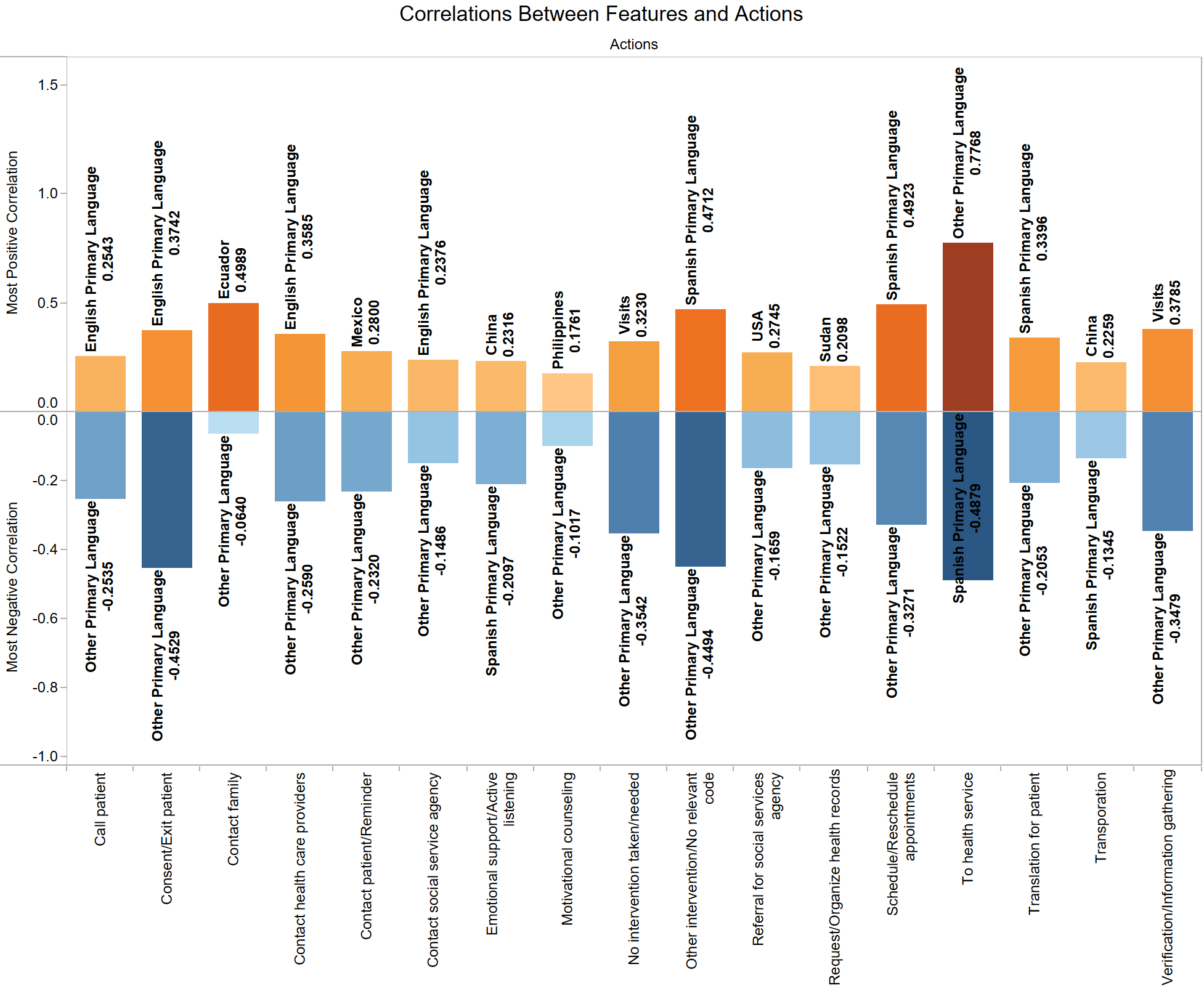
Appendix A



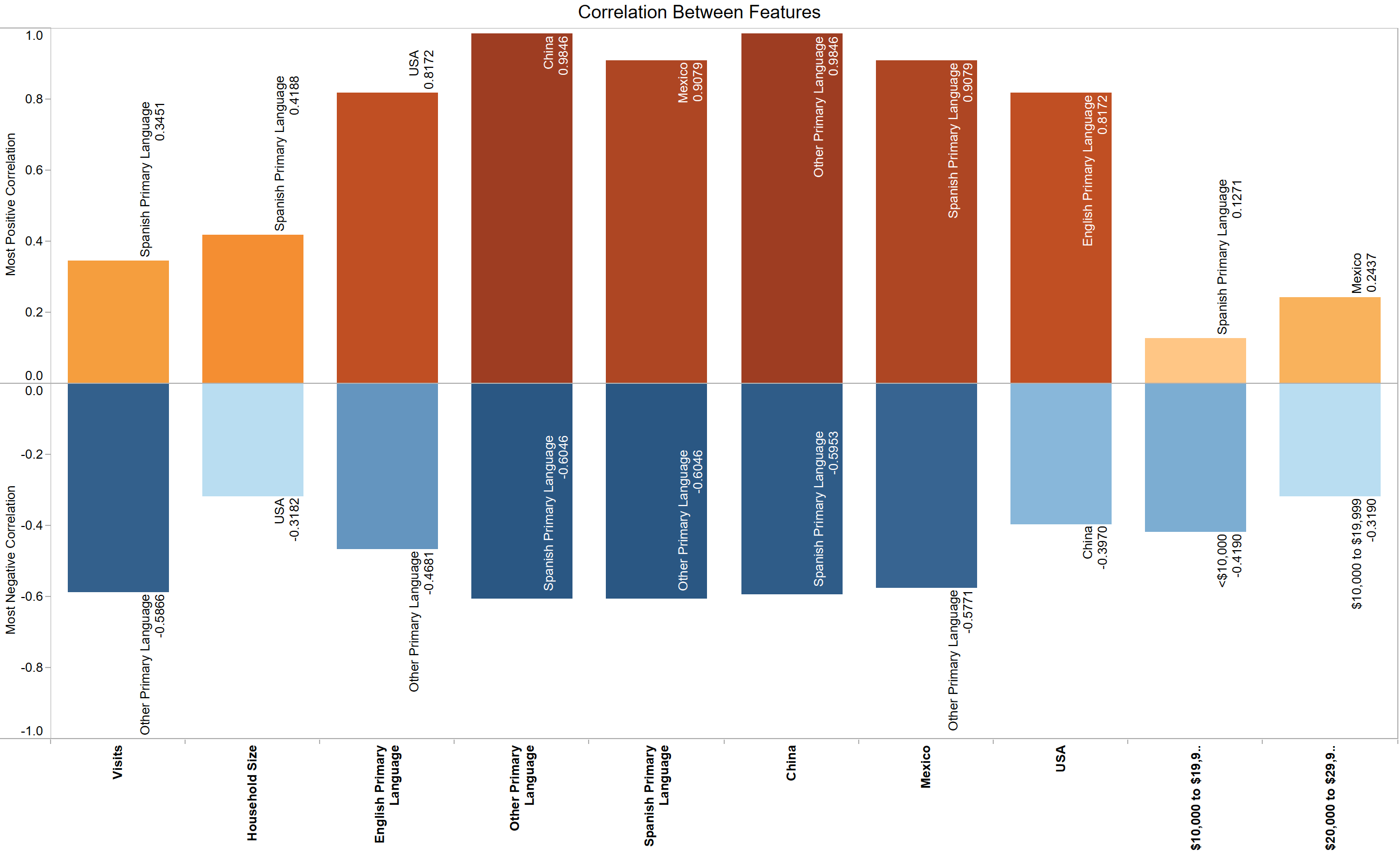


Appendix B





Appendix C



Appendix D

