Agency: Washington County

Project Title:

Strengthening Race-Ethnicity data through open-source development, machine learning, and regional collaboration

Project Description:

The use of data and information technology by local public health departments have increased due to the COVID-19 pandemic and raised expectations of timely disease surveillance, evaluation of complex health needs and outcomes, as well as meeting national accreditation requirements. This escalation of health data use requires skills in the field of health data science is beyond the training and job description of most epidemiologists and data analysts. The Research, Analytics, Informatics and Data (RAID) program within the Public Health Division provides the division's diverse research and data analysis needs.

At times, the inability to collect and analyze sufficient race-ethnicity data has crippled public health's ability to identify important health disparities early in emerging outbreaks and pandemics. This gap has demonstrated an explicit need for analysts and epidemiologists alike to understand how to maximize the utility of race-ethnicity data to advance health equity. The objective of this project is three-fold: to improve the quality and completeness of race-ethnicity demographic data collected in the county, develop best practices for using race-ethnicity data, and to increase transparency and sharing of code/methodology across jurisdictions in the state of Oregon.

In 2020, the Oregon legislature passed a law that requires health care providers to collect new standards for Race, Ethnicity, Language, and Disability (REALD) information at health care encounters related to COVID-19 and share this information with the Oregon Health Authority (OHA). OHA curated a REALD Implementation Guide that was used to create and align methods for categorizing and aggregating REALD data. However, additional detailed methods are needed to make the most use of the data. There is a need to better align these advanced methods across teams and data sources that collect REALD information. Our project will implement and compare text mining techniques that will be integrated into our code base. We will use pattern matching with REGEX to hard code REALD, manually validate results to create training and testing data sets, and then use different classification machine learning algorithms to automatize the process of obtaining the right REALD category for a person using the open text fields available. Some but not all the models we will implement are: Naïve Bayes, Neural Networks, Support Vector Machines, Binary Presence, Frequency Weights, and Random Forest.