FractureProof: An open source artificial intelligence process for precision public health

Andrew S. Cistola, MPH

Precision Public Health (PPH) involves the application of geographic data science techniques for small-area analysis to create targeted public health interventions. As a new area of health research, PPH strives to incorporate innovative tools in machine learning, artificial intelligence, and geographic information systems in a manner analogous to analyses of genomic data in Precision Medicine. Population health management (PHM) efforts will need to be driven by location-based analyses that incorporate social stratification and community infrastructure relevant to public health. While studies have identified that environmental factors are a significant predictor of clinical outcomes the full extent of ecological health impacts needs further investigation. PPH research has the potential to greatly benefit PHM efforts. However, there a number of obstacles in conducting PPH research.

When ecological data is collected, methods for identifying an actionable set (5-15) of possible factors have not been well established. Many common methods using probability based statistics do not have the ability to handle data from multiple geographies (Zip Code and County) with high dimensionality (2000+ variables) without significant limitations. Many new approaches in Artificial Intelligence that utilize “black-box” algorithms can predict outcomes very effectively, but do not provide interpretable results for understanding possible causal pathways. Similarly, many advanced algorithms suffer from “overfitting” or “noise” and may not provide informative results when deployed without proper processing. PPH studies will need to be validated by showing that public health efforts can be improved by results. This will require designing studies that can be replicated in various contexts and contain information relevant to practitioners. Using Artificial Intelligence in PPH research can be difficult for many entities that

lack the computational resources and technical expertise in order to translate information effectively.

FractureProof uses various algorithms for specific purposes in identifying important predictors that account for high variation while accounting for geography and eliminating redundant measures. A final list of selected features is evaluated using multiple regression modeling and receiver operator curves. This allows for researchers to search large datasets for possible causes, avoid type 1 errors familiar, and comfortably evaluate the significance of the results. FractureProof automates this process so that exploratory analyses can be conducted without prior hypotheses or feature engineering. The FracureProof process uses open source algorithms that have been widely deployed in many fields and have significant presence in scientific literature. Algorithms are taken from commonly used Python libraries including: scikit-learn, keras, statsmodels, PySal, numpy, scipy, pandas, and geopandas. Each of these libraries are available through the Anaconda Distribution, which provides users with the Python programming language, graphical user interface, easy to use software, and development environments without cost. FractureProof is deployed through GitHub under the MIT license and available as a software package for download to local devices. FractureProof does not require advanced computational resources and can be used on desktop devices common in enterprise settings.

FractureProof can be accessed at: https://github.com/andrewcistola/fracture-proof