Project Background:
Large-scale policies and programs, such as violent crime reduction initiatives and economic development projects, certainly have a substantial influence on health. However, current approaches to studying program or policy health effects have major limitations.

In traditional Public Health analyses the choice of study design and analysis approach for health effect assessments has been informed by general frameworks that discuss pros and cons of different designs and analyses. While these frameworks provide useful broad guidelines, a more rigorous quantitative assessment of which specific design and analysis combination is best to answer the scientific question of interest, given characteristics of the program and potential biases, would greatly improve study quality. Further, a rigorous assessment would involve considering many more design and analysis methods than in traditional analyses. Thus there is a critical need for a rigorous and cost efficient system that can be applied to assess the post-implementation effects of policies and programs on a wide range of health outcomes.

We aim to establish a new system to determine the health and health disparities impacts of policies and programs. The system will include a simulation generator to identify the optimal study design and analysis combination for a specific program or policy. Epidemiologists, public health professionals, government agencies and non-profits will be able to use the system plan studies and data analyses for the evaluation of policies and programs.

Users will provide information about their data and the quantities they aim to estimate. The simulator will conduct an array of statistical analyses, leveraging the wide range of analysis tools available in the statistical programming language R, to identify the best combination of study design and analysis for obtaining estimates of study effects. This is a computationally intensive task due to the wide range of potential study design and analysis combinations, the use of machine learning algorithms, model cross-validation, and bootstrapping to estimate variances.

AWS Solution:
First we will benchmark our simulation system and project the costs our system would incur across the range of EC2 and EMR possibilities from the C3 instance series. Additionally, using EMR, we will benchmark and project costs for a range of cluster sizes and instance types. Our test simulations on a local cluster take approximately 7.76s/core/iteration for our analysis tests, and 8.25s/core/iteration for our study design tests.

Second, we will conduct the first real-world tests of the system in order to choose the study design and analysis combinations that we will use to evaluate the health effects of programs such as California’s Public Safety Realignment (AB109) and a violence prevention program in Oakland, CA.

Third, we will invite a small group to pilot the system, conduct a small number of simulations using the system, and provide us with feedback.
Fourth, we will launch the beta version of our application to the public and invite users to provide further feedback. We will advertise our system by presenting and conducting training events at academic conferences and online through video and interactive web-based tutorials, publishing papers demonstrating the use of our system on real-world problems.

Usage requirements:

- We will test our system on a variety of different instance types using EC2 and EMR, in the US-West (Northern California) region:
  - EC2:
    - c3.2xlarge: ($0.42/hr)*(~280 hr/test)*(10 tests) = $1,176
    - c3.4xlarge: ($0.84/hr)*(~140 hr/test)*(10 tests) = $1,176
    - c3.8xlarge: ($1.68/hr)*(~70 hr/test)*(10 tests) = $1,176
    - Total EC2: $3,528
  - EMR: We will conduct similar simulations using EMR, along with varying the number of instances (~5-50), totaling ~$3500.00.
  - Assuming c3.8xLarge EC2 for post-benchmark activities:
    - 3 instances at 24 hr/week c3.8xLarge = 530.26/mo ($12726.24 over 2y)