**Problem Statement**

Delays in supply of commodities result in extra costs in terms of storage, coordination and most importantly lost lives in the case of HIV medicines. This study will use publicly available supply chain data to determine the most important factors in predicting whether HIV drugs are delivered on time or not. It will then use these factors to predict how long these delays will be, thus allowing HIV program managers to know **when and which products are likely to be delayed**, as well as **the extent of the delay** so they can take mitigating action to save lives and avoid additional supply chain costs.

**Solution Statement**

As a proposed solution, this study will first explore **classification machine learning algorithms** to determine whether a particular product was delayed or not. It will then use **regression analysis** to predict the length of the delay using the subset of the data which the classification predicts will be delayed. This will maximize the utility of the complete model since it follows the natural decision-making process – a program manager would normally care about the products that will be delivered late and within those, focus on the ones that will likely have the longest delays first, thus allowing them to prioritize supply chain/logistics management and solve the biggest problems first.

To select best model, both the classification and regression versions of these models will be explored evaluated against the benchmarks (see **“Benchmarks”** section below): i) Random-Forests ii) XGBoost iii) Support-Vector Machines (SVM) and iv) Recurrent Neural Networks (RNNs). Random-Forests and XGBoost are proven **high-performing ensemble** algorithms which can do **automatic feature extraction** while SVMs perform very well with **high-dimensional data** and can **detect non-linear relationships** if the right kernel is used. Finally, RNNs are useful for high-dimensional time-series data. Please see **“Project Design”** section for workflow and overview of algorithms. The above advantages of these algorithms are well-suited to the selected dataset which has several categorical columns which will increase dimensionality and potentially be non-linearly related to the target variable after data transformation. Finally, the data is well-suited for this overall approach since our target variables is well-defined on the data i.e. it can be determined by data on scheduled versus actual delivery dates, thus allowing us to quantify and measure the problem and solution. The study results will be applicable to future instances of supply chain orders, and thus it is applicable to future occurrences of similar supply chain data observations.