More than **36.7 million1** people in the world were living with HIV in 2016 and every year, about **1 million** people worldwide die from AIDS-related causes. While this death rate has decreased significantly (by 38%) since 2001 and continues to decline, about **1.8 million** people also became newly infected in 2016 alone. The epidemic disproportionately affects low income countries in Eastern and Southern Africa where women, adolescents and key populations like female sex workers and LGBTQ individuals are the most affected groups. There is currently no cure or vaccine for HIV and while several prevention methods exist, their efficacy is reduced by several factors, including economic and psycho-social factors. Fortunately, it has been shown that treatment can not only prolong life but also prevent the spread of HIV as it lowers the viral load of people living with HIV to a non-infectious level. However, of the **36.7 million** people living with HIV in 2016, only **19.5 million** were receiving this life-saving treatment.

The President’s Emergency Plan for AIDS Relief (PEPFAR), a US government program is a key player in the procurement of drugs, testing and laboratory kits for HIV. One of its agencies spends more than **$9.5 Billion** per year on procurement of essential medicines to fight HIV/AIDS around the world. It is critical that these procurements arrive on time and in full to meet the needs of People Living with HIV (PLHIV) around the world, however this is not always the case. In fact, recent changes have resulted in drastic declines in supply chain performance as shown in the chart above. Thus, knowing whether or not HIV drugs are delivered on time and how long potential delays will be is very important. This study will use publicly available data from PEPFAR over the years 2006-2015 to determine the factors influencing timeliness of pharmaceutical deliveries as well as use these factors to develop a model that can predict if and by how long a particular HIV commodity will be delayed in delivery. While more and more supply chain analysis has begun to incorporate machine learning, it is especially aimed at demand forecasting as opposed to predicting the lead-time directly. However, the approaches taken in some academic studies2 e.g. SVMs and RNNs have shown great promise. Similar problems like predicting flight delays3 and improving flight efficiency have also been solved using machine-learning.



Timeliness of HIV procurement is critical to the efficiency and impact of the program in saving lives, controlling and eventually eliminating HIV. 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.

This study used a combined model which uses **classification machine learning algorithms** to predict whether a particular product is delayed or not and then use **regression analysis** to predict the length of the delay using the subset of the data which the classification predicted will be delayed. This will maximize the utility of the complete model since it follows the natural decision-making process – a supply chain 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 the best model, both the classification and regression versions of the following models will be explored evaluated against predetermined benchmarks of Random Forest model with default parameters in SciKit-Learn : i) ExtraTrees ii) XGBoost iii) Support-Vector Machines (SVM) and iv) Multi-Layer Percepton (MLP). Random-Forests, ExtraTrees 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, MLPs are useful for high-dimensional time-series data. The above advantages of these algorithms are well-suited to the selected dataset which has several categorical columns that 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. delay occurrences and duration can be determined by data on scheduled versus actual delivery dates, allowing clear quantification and measurement of the problem and solution. This study’s 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 and useful for planning purposes.