[still a lot to add in]

[ref about drug misuse and new resistant strains of bacteria/virus]

Infectious diseases are responsible for 30% of all deaths in Africa. [WHO but check source]

The World Health Organization (WHO) describes these as “caused by pathogenic microorganisms, such as bacteria, viruses, parasites or fungi; the diseases can be spread, directly or indirectly, from one person to another [WHO website].

These diseases, among which HIV, tuberculosis, Malaria, Hepatitis and neglected tropical diseases (NTD) have caused 4.3 million deaths in 2016, down from 5.3 million in 2000. Still, the proportion of people dying from low-income countries is much greater than the rest, with reports showing that Africa and South East Asia are the most impacted [ WHO report 2019 infectious diseases]

These diseases are complex and so are their evolution and spread. However there are some factors which put certain populations or geographical zones more at risk than others.

Notably, doctor / clinician availability is not the same in all places: low income countries tend to have less than 10 doctors per 10,000 inhabitants while high-income countries don’t (respectively 90%,5%). Access to medical care and facilities, the number of clinicians in the population, the state of medical infrastructures (as well as transport infrastructure which allows to get to these facilities) are very different across the globe. Moreover, some clinicians may not be as trained or not trained for as long as in wealthier countries.

Some other factors are climate-related, although humidity, rainfall and temperature are correlated both positively and negatively with different infectious diseases. However, some diseases are mostly transmitted through insects such as mosquitoes, and this species benefits greatly from humid seasons to reproduce. Climate change will most likely push climate variations to new extremes and could potentially increase the risk for propagation of infectious diseases.

There is also a strong economic factor that comes into play. With more financial resources, diseases can be more adequately treated, not just because health infrastructures are more present and up-to-date but also thanks to a higher number of health workers and to the ability for both the state and the individual to pay for more expensive – and often more efficient – treatments.

Another key aspect to infectious diseases is its transmission. In fast-growing countries, there are both very rural places where contact to animals is frequent and health and safety procedures with regards to animal contact are minimal. Animals, and insects which accompany them, being one of the main factors of transmission of diseases to humans, this behavior can lead to greater risks of outbreaks. These same countries also possess fast growing cities and urban centers with high population densities. This is the case for most countries, yet wealthier countries often possess more health infrastructures. If infectious diseases start to spread in these centers, they might be harder to stop.

Another key aspect of the problem at hands is epidemics and pandemics. The Covid-19 outbreak is a great example of how impactful on the short and long term these disease outbreaks can be. Recent history has shown that pandemics and epidemics thrive in countries that are less wealthy. The Spanish flu epidemic has affected India very strongly in terms of fatalities. However, Covid-19 seemingly has affected wealthier countries more. This can be in part explained by different government responses and co-morbidities such as age and obesity which are less frequent in less wealthy countries. Although countries in Africa and Asia have response and watch/sentinel protocols put in place, the lack of resources still means they would be disproportionately affected. Detecting patients in real time or at least soon enough is very difficult, especially if there are not enough resources at your disposition.

The use of drugs if the last aspect of infectious disease we will cover. It has been shown that many antibiotics are used on patients that are not infected, which leads to both a poor attribution of often scarce resources and a rampant rise in antibiotic resistance in microbes. This is exacerbated by the fact that doctors tend to err on the side of caution, meaning that if an antibiotic is available for a patient, even if the patient has a very slight chance of having a disease, it will be prescribed. Drug resistance has become a clear issue in the 21st century and leads to most medications becoming ineffective. When these become useless against a disease, the alternatives are few and more expensive. Moreover, alternatives are not infinite and the pool of antibiotics to choose from shrinks considerably with time.

These problems are currently addressed in part by the WHO. A guideline (IMCI) has been written and put in place in order to give some information to clinicians on possible diagnostics. Since these guidelines are static, they are constantly outdated. Moreover, they are very broad and general, and since disease usually have a specific geographic and climate based context, are often not optimal.

IMCI guidelines fall in the CDA category, along with many different and new algorithms.

In an effort to palliate these issues, electric CDAs have been put in place. However, most of these are still static and not geography-specific.

While the use of e-CDAs is a step forward, there is ample room for improvement with the use of ML, data science and data collection. A new technique, e-POCT has been put forward, with the general idea being to collect data during point-of-care tests and to use it to gather information not only on a patient but on a population and its evolution with time. This provides much needed granularity of population analyses in both the geographic and the temporal dimensions.

The goal of the CliniColab project is to bring specificity to the issue being addressed. For a specific disease, any clinician can gather basic knowledge on how to act from IMCI guidelines, and some context knowledge from years of expertise in their own clinics or from their collaborators. The idea of CliniColab is to allow clinicians to gain more insight on their patients and the context that is specific to their location.

The overarching idea is to use data science to address clinicians’ needs. Machine learning algorithms will be used to predict diagnostics at a certain threshold – based on the accuracy of other algorithms already put in place, or based on the resource consumption of other algorithms – or even improve overall prediction accuracy. A parallel goal is to give insight to the end-user on ways to improve the accuracy. To this end, uncertainty needs to be addressed.

There are many different factors of uncertainty in the diagnostic pipeline.

First, there is uncertainty in the data collected. Many values are often missing, values are not normalized or standardized across clinics or regions, some features useful to the ML prediction are disregarded when conducting point-of-care testing, etc… Data preprocessing also adds a potential source of errors. There is also general uncertainty in the data itself – meaning the process that is being measured, which is called aleatoric uncertainty. While uncertainty in the data collection process can be improved, aleatoric uncertainty is inherent to the data itself and is informative of the differences in the data. In general, data is dirty and data scientists have to deal with that.

Second, there is uncertainty that is specific to the model being used which does not stem from the data, which is called epistemic uncertainty. This uncertainty should be reduced as much as possible.

Third there is uncertainty in the results. This uncertainty can be mitigated using confidence intervals on the statistics outputted, or using interpretation and explanation methods that are suited to the model and data type. Confidence intervals are paramount in domains such as medical diagnosis, since most classification predictions are probability-based and should be considered with their associated probability. Another approach to giving information to the end-user about the probabilistic side of predictions is to use Bayes-neural nets.

It is also vital to be able to explain all or most results. For example, a result that seems to be very different from the rest of the predictions, and that is unexpected, can be analyzed using interpretability tools to gain insight on what in the data or in the model caused this prediction. Understanding edge cases and acknowledging them leads to a better understanding of the model. It is important not to blindly trust the model predictions, but to understand in what it succeeds and in what it fails, even if it does admirably or poorly.

Once these uncertainties are lifted in part or in whole, conclusions can be drawn from the ML algorithms’ outputs. These conclusions can give insight on the population being analyzed or can help detect trends and possible indicators of diseases.

The project is focused on interpretability of the algorithmic tools used.

Understanding what causes outputs to be predicted can help with data collection: if a user knows that having data points with a certain feature helps the prediction, data collection of this particular feature can be insisted upon. In general, having an understanding of how the model behaves yields powerful insights on the value of the features currently available. Moreover since ML algorithms do not behave as clinicians do, the importance or lack thereof of some features might not be intuitive – most people might require to analyze the ML model’s behavior in detail before understanding from which features it benefits more. Feature engineering also benefits from this knowledge.

Focusing on interpretation is paramount in ML applied to medicine. Since humans are ultimately responsible for decision making, it is important to give both mathematical results and insights on why these results happen. This is a key component of the trust-building process for clinicians.

In the context of our platform the interpretation tools we use should be model-agnostic, meaning that they can be used with any type of model, since the platform offers a range of model choices which differ in their structures.

Overall, interpretation tools can be split into two categories; those that focus on the entirety of the model’s outputs, and those that focus on single instances or regions around an instance.

One of the interpretation methods that is focused on the dataset as a whole is permutation importance. It is a measure of the impact that different features have on the predictions. The process happens after the model has been trained. The chosen feature is randomly shuffled in the test set, while all other features remain the same, and the ML is tested on this newly partly shuffled dataset. Multiple shuffles occur, and the results of the ML model using shuffled version and normal non-shuffled versions are compared. This technique offers a way of comparing the feature’s importance to a prediction versus randomized versions of itself. If the model relied on this feature for predictions, then there should be a drop in accuracy with the randomized version. Most techniques use more than one shuffle for robustness, therefore the importance score is accompanied by a variance measure. Positive values show that the feature is useful for predictions, with higher values indicating higher importance. Negative values can happen if the overall importance is very close to zero (Feature is not useful to the model prediction) but its shuffled versions are randomly a little better.

Another interpretation tool are partial dependence plots. This tool considers a single feature’s impact on the prediction while varying another single feature’s value. This method is flawed in many cases, but it can give an intuition on the relationship between these two values and the way they interact with respect to the prediction.

SHAP values are an upgrade of the PDP. These explain individual predictions but can be then aggregated over the whole dataset to gain general insights. The idea is to compare the feature’s importance with different values with respect to a baseline. SHAP values can be used on dependence contribution plots.   
SHAP values allow to measure the feature’s sole contribution – disregarding dependence effects caused by correlation to other features or other interactions.

Another method is called LIME, which explains individual predictions by defining an interpretable model that approximates the model’s behavior well enough in the neighborhood of that data instance. When looking at the simpler model, it can be more easily interpreted (e.g. consider a linear model with few coefficients).

In order to deal with output uncertainty, CI can be used. Since the distribution of the input data is usually unknown, it is advised to use a bootstrap resampling method: samples are drawn with replacement from the dataset, and the parameter of interest is estimated on each new sampled dataset. When repeating this process a number of times, a robust estimate of the population’s parameter can be estimated from this sampling method.

Concerning the results, there are also very direct ways to interpret them. It is important to remember that the results need to be analyzed before delving into their interpretation and the way the model predicted them. There are several metrics that are of use in the medical field, accuracy: (TP+TN)/(TP+FP+TN+FN); precision: how much the model is right when it says it is: TP/(TP+FP); recall (sensitivity): how much right did the model miss: TP/(TP+FN); specificity: TN/(TN+FP); F1 score: harmonic mean of precision and recall; PR curve: precision vs recall; ROC curve: receiver operating characteristic: true positive rate vs. false positive rate so recall vs (1-specificity); AUC is the are under the curve: AUROC.

In order to reduce epistemic uncertainty, it is important to conduct cross-validations and use a hold-out test set.