Capstone 1: The Fight Against Malaria
…An effort at predicting the impact of malaria

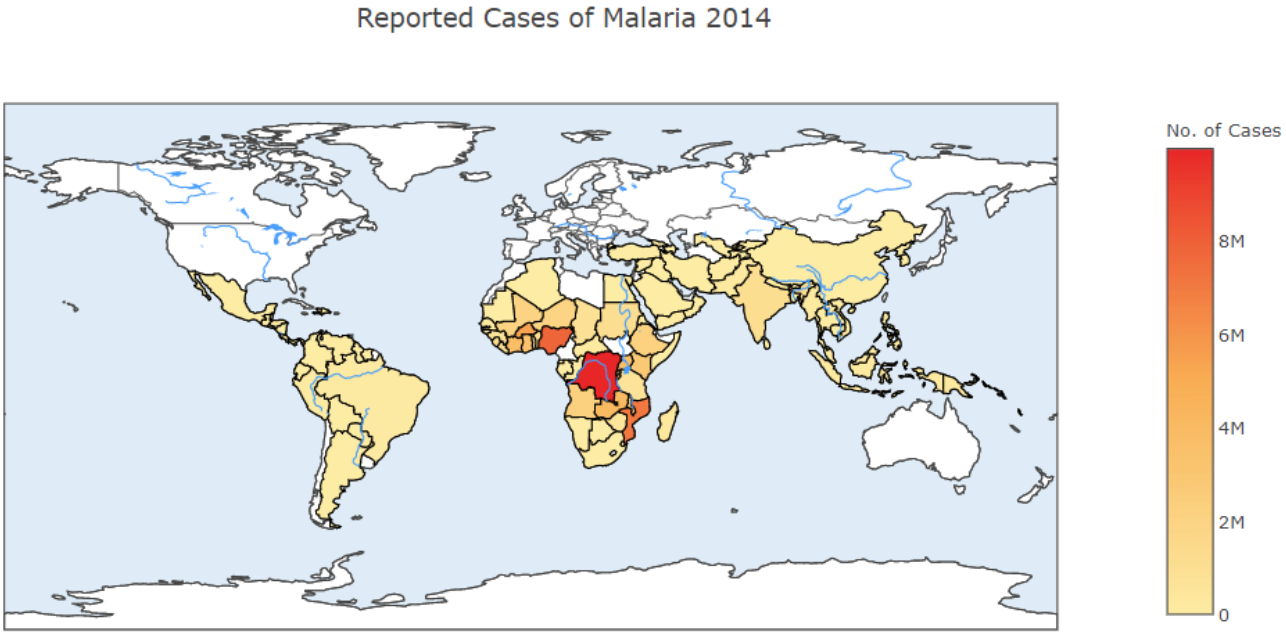
Milestone Report 1

Introduction and Goal

Malaria, though rare in the USA (most cases present in recent travelers), is still a prevailing pestilence that claims hundreds of thousands of lives across the world. The WHO estimates that in 2015 there were 214 million new cases of malaria resulting in 438,000 deaths. Geographically 40% of the world's population lives in regions with a risk of malaria. From an economic perspective malaria costs Africa alone an estimated $12 billion per year in GDP. This is all in the backdrop of modern public health and medical science having the ability to greatly curtail this impact as we have done in many areas already but despite efforts of many organizations we have failed to do so to any degree of completeness.

In the late 19th century malaria was endemic in shaded regions,
***Source:** Reiter, 2001.*

However, resources are not unlimited and must be directed to the area's most in need and/or where they would be most effective. Further, we must be able to objectively determine if and to what degree these efforts are succeeding. The core goal of this project is to address some of these factors and to raise awareness of the general impact that malaria still has on the developing world.



There are two primary goals of this project. One, to determine if efforts to distribute insect repellant nets are effective. Two, using various environmental and social parameters known to affect the spread of Malaria I will apply machine learning methods to predict the incidence of Malaria in a given country and thus provide an estimate of the burden of disease and the ability for relief efforts to be more accurately directed.

Though not entirely necessary a basic understanding of the life-cycle and pathophysiology of Malaria will help with contextual understanding of this project. A quick summary of relevant information can be found in Appendix I. **N.B.** – this report utilized figures and information both used in this project and from scientific literature. This was done both for informational purposes (I do not have access to the same data sources) and as a sanity check to ensure that at least the general trends I was finding matched those seen in work by those at the WHO etc. Anything labeled and sited is adapted from outside sources.

Data Acquisition

Given that the data set used for this project comes from a variety of sources and files I will be breaking it down by the features that are present in the working data frame.

The original premise and starting data were obtained from a Kaggle competition here. This was then supplemented by data taken directly from the World Health Organization (WHO), The World Bank, and GapMinder. Even amongst the major organizations that track such data there are, at time, major discrepancies in their estimates. This can be attributed to variation in methodology.

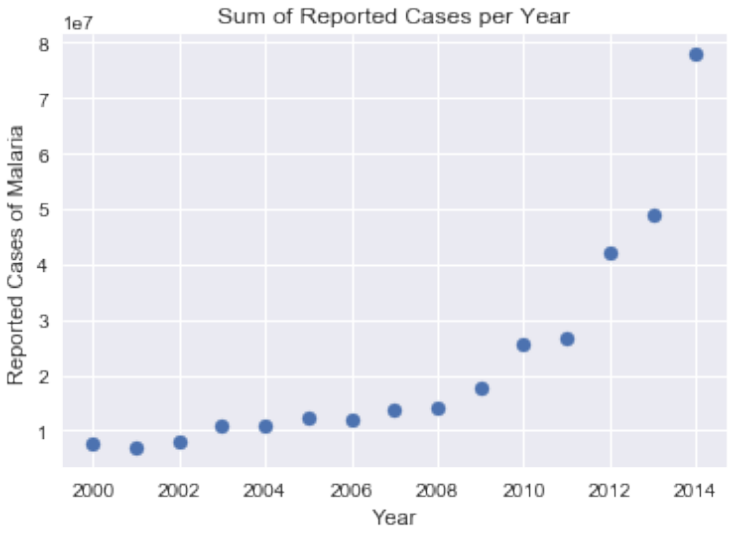
So, given that this data has originated from differing sources and that some has been averaged it will not be as precise as some tasks would require. Herein I am looking for trends and general associations. A small amount of variation is acceptable to this task. Further as a baseline obtaining accurate and reliable information about the incidence and impact of Malaria has plagued investigators for a very long time.

Even when information is available it is often brought into question. This is due in no small part to the fact that record keeping and transmittal of data from indigenous regions where Malaria has the highest impact to those doing the research is not very streamlined to say the least. On top of this in the field Malaria cannot always be differentiated from other tropical diseases thus causing inaccurate or missed diagnoses.

Data Wrangling and EDA

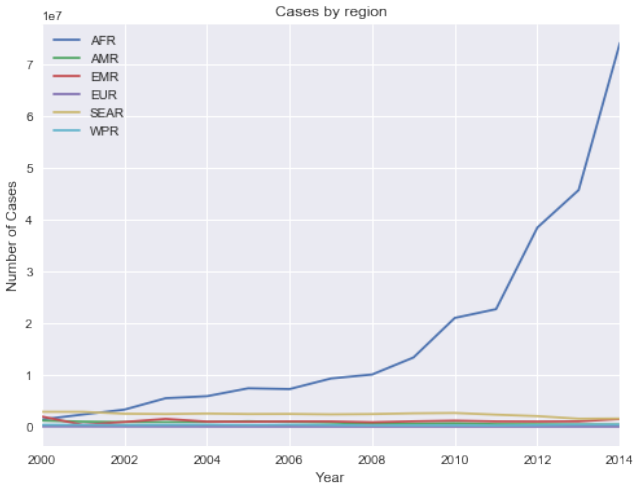
As previously discussed the data is being combined from several CSV files. This was done on a basis of county and year. With the result of each line of the final data frame representing a year time point in a particular county. Given inconsistencies in time frames and reporting there were many instances of missing values in one or more years for a given country. Data points with missing values were dropped from the analysis. When looking into some of the individual data features there were some interesting things to be found.

Reported cases of disease:



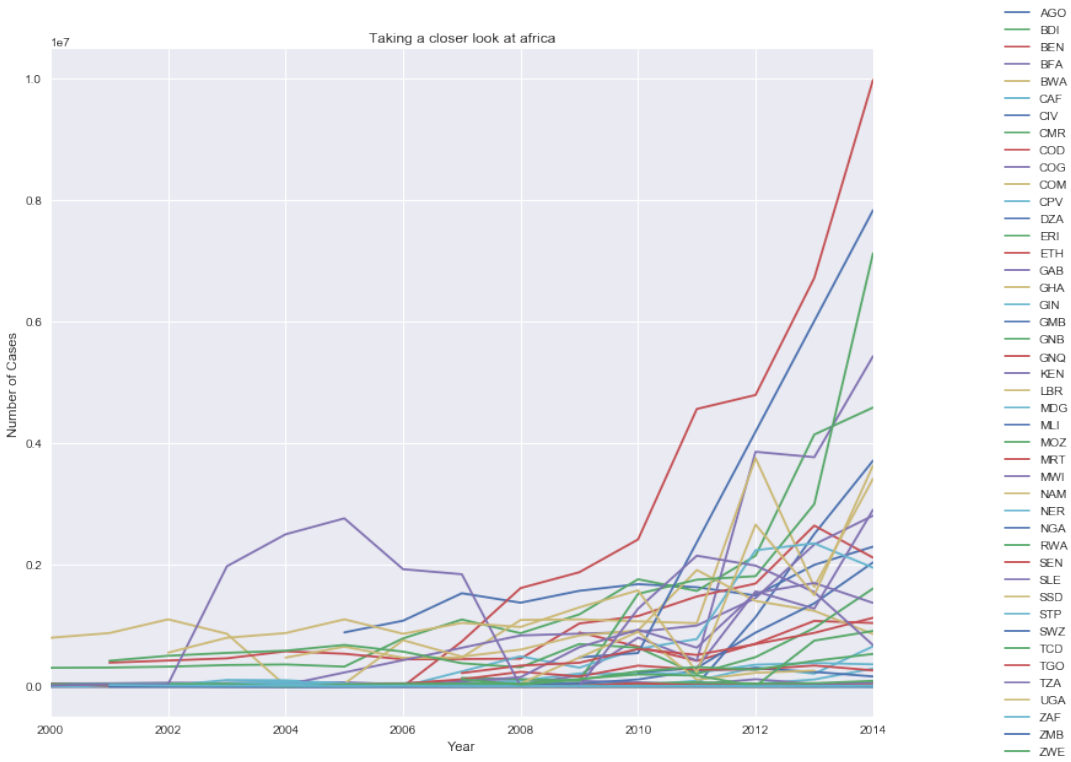
Here we can see a steady climb of reported cases of malaria. Now is this an objective climb in numbers or are we getting better at gathering accurate numbers? Either way this trend is mirrored by patterns seen in the WHO and IHME data. One sanity check complete.

Looking deeper we can see:



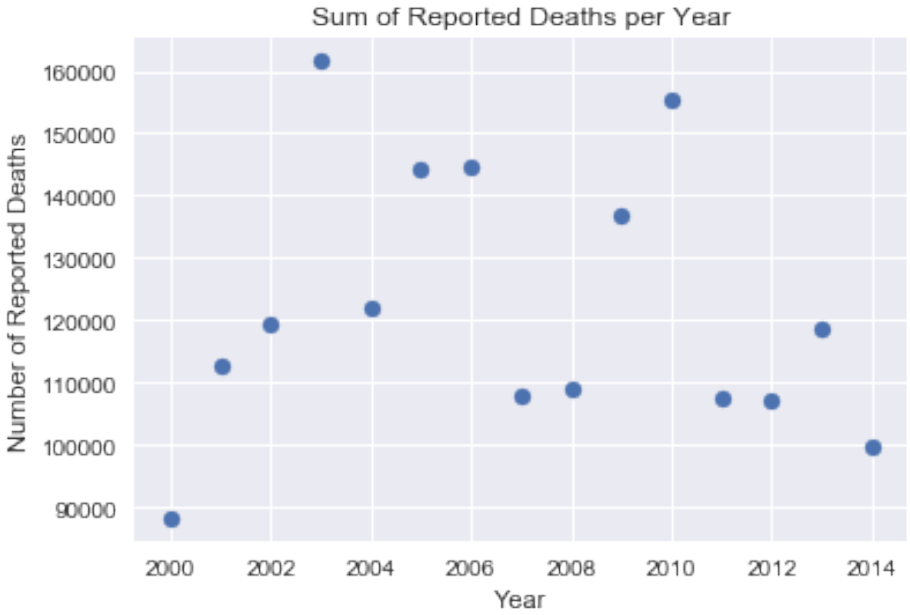
From this it is very apparent that the overall trend and a great deal of the impact attributable to Africa.

Zooming to look exclusively at Africa:



We can observe one or two countries that have consistently higher case number but not as great difference as seen between Africa as a whole and the rest of the world.

What about mortality?



This one is a bit more variable, but it does seem that we have an overall decrease in number of deaths in recent years. This data overall matches with the below from ourworldindata.org.

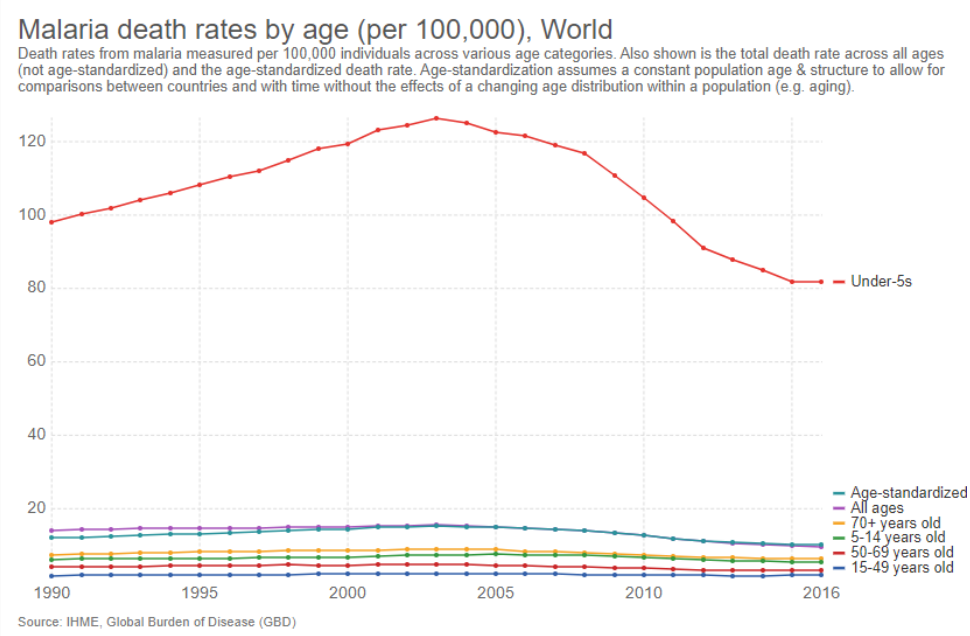
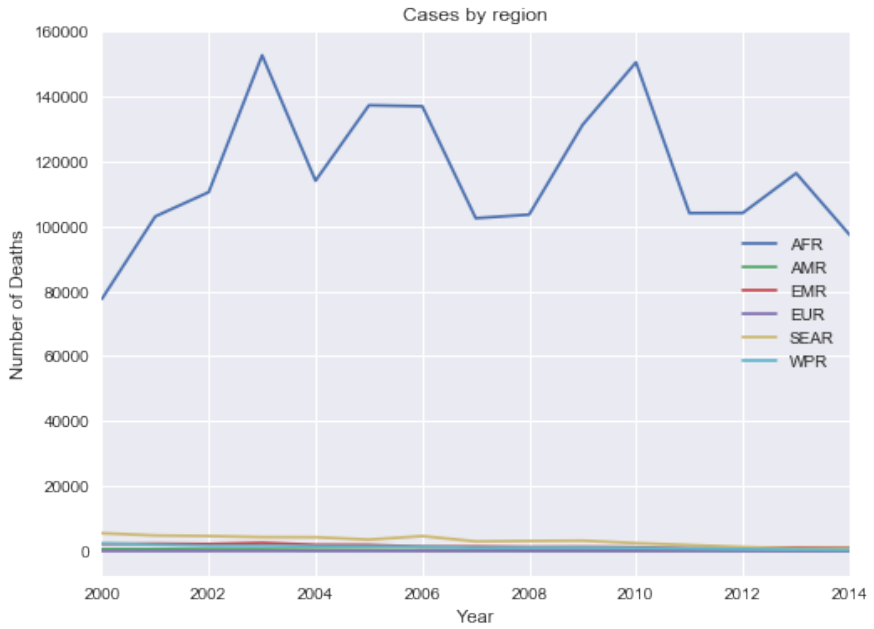


Figure 1 Malaria deaths by age group (Adapted from Roser and Richie)

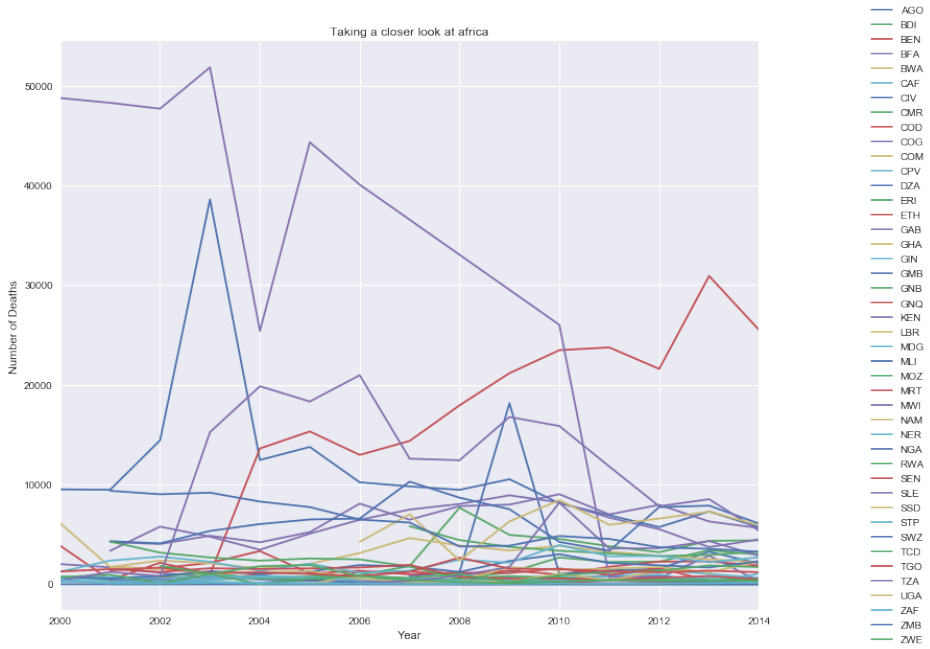
A few things about this chart: one, it matches in overall flow of the project data with a relative spike in numbers in 2003. Two, it appears more smoothed due to time frame and perspective differences. Three, the data for this project comes from the WHO the data used for the chart came from the International institute of Health Metrics and Evaluation (IHME). Four, here we are also able to see the proportion of deaths by age group. Notice that children under 5 are suffering a much higher death toll than the other groups. Sanity check two down.

Deaths by region:



Here too Africa seems to be the principal contributor to the overall trend.

Looking at Africa by country:



We have 3 possibly 4 countries with being the hardest hit with the remainder bunching together a lower death numbers.

From a more quantitative point of view:

country_code	year	reported_cases
COD	2014	9968983.0
NGA	2014	7826954.0
MOZ	2014	7117648.0
BFA	2014	5428655.0
BDI	2014	4585273.0

We have a mismatch between morbidity and mortality i.e. the number of cases are not necessarily determining the burden of death on any given country. This is not particularly surprising, but it does emphasize that there are factors at work that predispose particular countries to have higher incidences of disease and likely differing ones that predispose to higher death rates within the disease population.

Some perspective can be gained here:

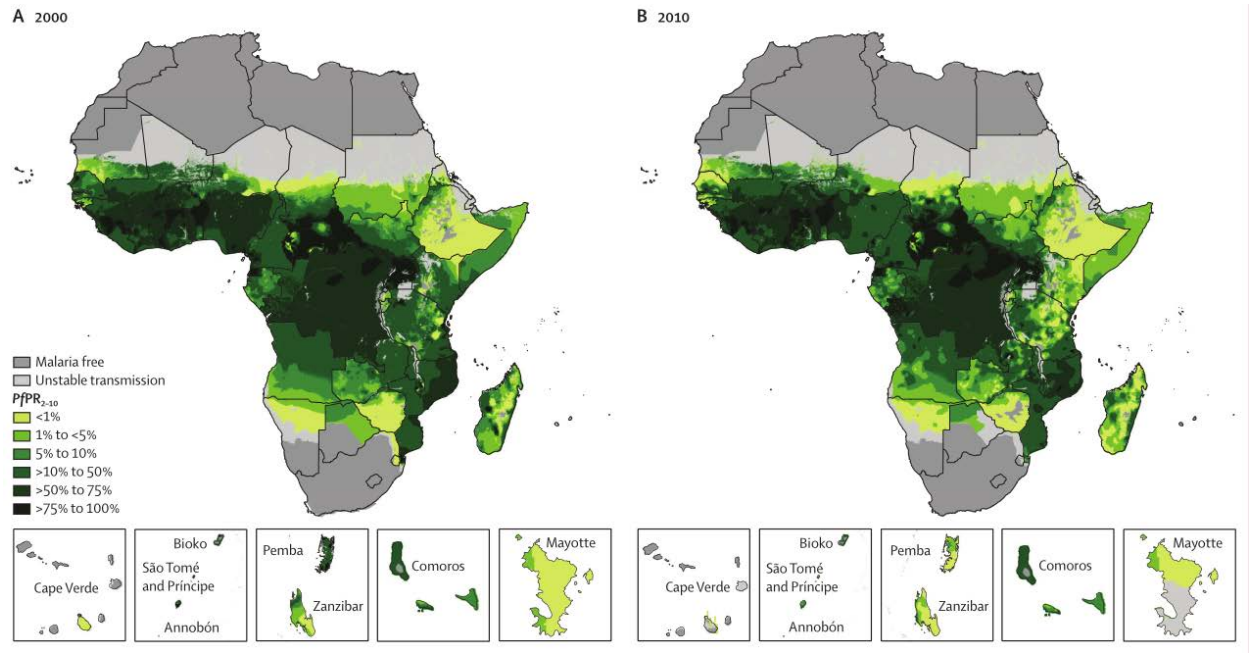


Figure 2 Adapted from (***)

The above heatmap of malarial prevalence has a lot going on but the core point is that Africa particularly the central regions does in fact have a highly pronounced level of malaria. Also given that the left side depicts the levels for 2000 and the right for 2010 with little change in density it would appear to be a very strong hold on those regions.

The Combined data:

The year range for some of the data goes back several decades but from examination the time frame with the highest density of data points turned out to be from 2000 to 2014. The resulting data looked as follows:

```
Number of years:  15
Years from:  2000 - 2014
Number of countries:  90
year                False
country_code        False
reported_cases      False
region_x            False
reported_deaths     False
region_y            False
rainfall            False
temperature         False
population          False
country_name_x      False
percent_agg         False
percent_urb         False
country_name_y      False
```

```
gdp_per_cap        False
country_name       False
pop_density        False
incidence          False
dtype: bool
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1075 entries, 2 to 1390
Data columns (total 17 columns):
year               1075 non-null object
country_code       1075 non-null object
reported_cases     1075 non-null float64
region_x           1075 non-null object
reported_deaths    1075 non-null float64
region_y           1075 non-null object
rainfall           1075 non-null float64
temperature        1075 non-null float64
population         1075 non-null float64
country_name_x     1075 non-null object
percent_agg        1075 non-null float64
percent_urb        1075 non-null float64
country_name_y     1075 non-null object
gdp_per_cap        1075 non-null float64
country_name       1075 non-null object
pop_density        1075 non-null float64
incidence          1075 non-null float64
dtypes: float64(10), object(7)
```

Note: this will be the data to which machine learning techniques are applied in the next step of this project. As such it will be examined more thoroughly in the next report.

All of the following analysis was conducted from the following data frame:

From the original Kaggle data a file with data detailing the distribution of insect repellant nets was provided. The data within when combined with the others that I had obtained restricted the number of complete data to an unacceptably low number. Therefore, I decided to create a second data frame to examine the impact of net distribution and insecticide resistant mosquitoes independently of other factors. Thus, allowing me to still look into some of the original goals of this project. The resulting data frame:

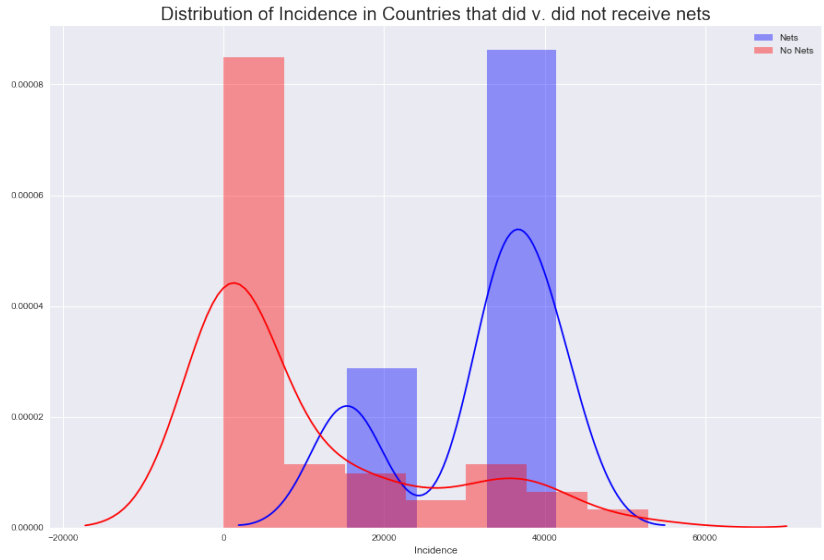
```
Number of years:  16
Years from:  2000 - 2015
Number of countries:  98
year                 False
country_code         False
tx_resistance        False
tx_resistance_int    False
number_nets          False
country_name_x        True
country_name_y        True
reported_deaths       True
incidence             True
```

```
dtype: bool
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2169 entries, 0 to 2168
Data columns (total 9 columns):
year                 2169 non-null object
country_code         2169 non-null object
tx_resistance        2169 non-null object
tx_resistance_int    2169 non-null float64
number_nets          2169 non-null float64
country_name_x       241 non-null object
country_name_y       1409 non-null object
reported_deaths      1409 non-null float64
incidence            1409 non-null float64
dtypes: float64(4), object(5)
```

Does the distribution of insecticide resistant nets have an impact on the burden of malaria?

As a marker for the impact of malaria I will be using incidence as reported by the IHME. Incidence is a good general marker for the occurrence of disease in a population at a given time. More information on the epidemiological use of incidence can be found here.

Using this variable, I conducted a hypothesis test for a difference in the mean and median incidence of those countries that received nets and those that did not.



Null: There is no difference in the median incidence of the two population
Alternate: There is a difference between the median incidence of the two populations

Test For median:

stat, p, med, tbl = stats.median_test(nets['incidence'], no_nets['incidence'])

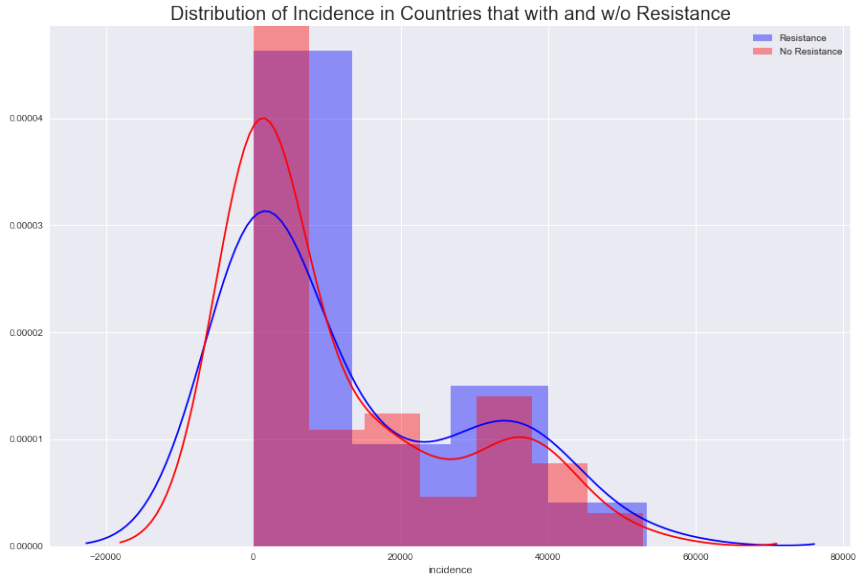
```
(2.4361033352473918, 0.1185, 1413.5992744999999, array([[ 4, 38],[ 0, 43]]
, dtype=int64))
```

Test for mean:
```
Ttest_indResult(statistic=2.9960178221521483, pvalue=0.0036)
```

From the p-value of the median test we cannot reject the null hypothesis, however by the p-value of the mean test we can reject the null. Thus we have a difference in the mean but not the median of the two populations.

Does the presence of insecticide resistant mosquitoes contribute to the burden of malaria?

Here I again used incidence. I conducted a hypothesis test for a difference in the median PP of malaria in countries that had documented resistance and those that did not. This was done with out consideration for net distribution. We will look at that in a moment. It is important to look at this as a general factor as insecticide is commonly used as a general preventative measure regardless of net distribution.



Null: There is no difference in the median incidence of the two population
Alternate: There is a difference between the median incidence of the two populations

Test for median:
stat, p, med, tbl = stats.median_test(resistance['incidence'], no_resistance['incidence'])

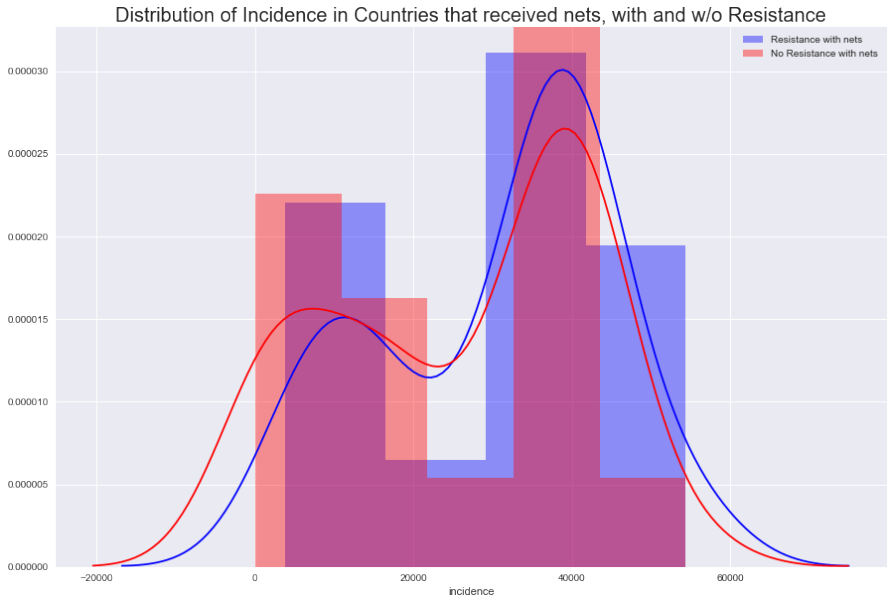
```
(0.0, 1.0, 1550.2330852499999, array([[28, 42],[27, 43]], dtype=int64))
```

Test for mean:

```
Ttest_indResult(statistic=0.55397101453946895, pvalue=0.5804
```

Here we see neither a difference of the median or the mean.

Now, looking for a difference in the countries that have reported resistance against those that do not **ONLY** in those countries that have received nets.



Null: There is no difference in the median incidence of the two population
Alternate: There is a difference between the median incidence of the two populations

Test for median:
stat, p, med, tbl =
stats.median_test(resistance_with_nets['incidence'],no_resistance_with_nets['incidence'])

```
(0.39669134132425848, 0.5288, 35819.00935, array([[29, 42],
    [32, 60]], dtype=int64))
```

Test for the mean:

```
Ttest_indResult(statistic=1.5916413121357946, pvalue=0.1134)
```

Once again, we see no difference in mean or median amongst the two sets of countries.

This then leads us to the question of how much of a difference do the nets and/or the presence of insecticide resistance make.

On a practical basis I would not recommend presenting a statement saying that the nets are ineffective. Some factors to consider the actual use of nets and adequate usage of those nets. Further, we must take into account our inability to truly track where and when resistant mosquito species will be. There is also the fact that people in malaria endemic areas spend a great deal of the day in mosquito infested areas e.g. agricultural work.

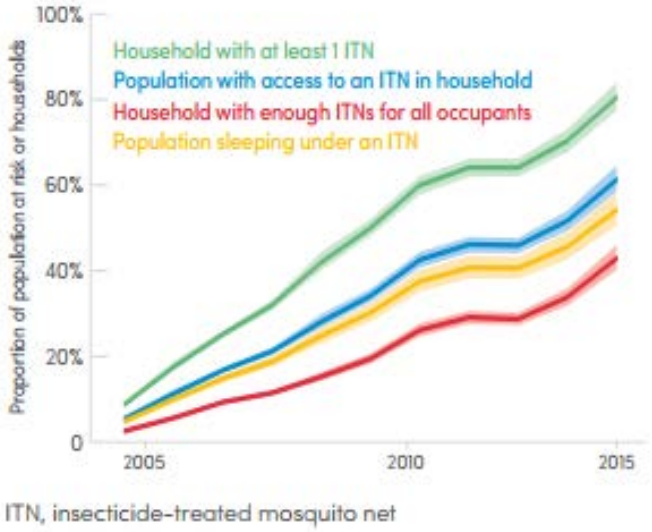


Figure 3 Some data on the availability and usage of mosquito nets, Adapted from the WHO 2016 Report on Malaria.

From the above figure we can see that there seems to be a goodly amount of availability of the nets but not enough for the entire house hold. A slightly more promising trend is that the amount of people actually using nets is only slightly below those with access to the nets.

Conclusions and Limitations

Given that the data is at a yearly and country level this limits the resolution of any findings from this investigation. Many researchers use a global perspective for discussing the impact of Malaria e.g. cost and death toll. However, many of those who attempt to accurately track and/or make predictions about Malaria related factors have generally used a country level approach.

There in they use similar data and methods as I have here, but do so by the month or even week on a time scale and from a county or city level on a geographic scale. This obviously increases usefulness and timeliness of results. One of the more interesting models is endeavoring to use satellite image data of weather patterns to give a several week warning of malaria outbreaks.

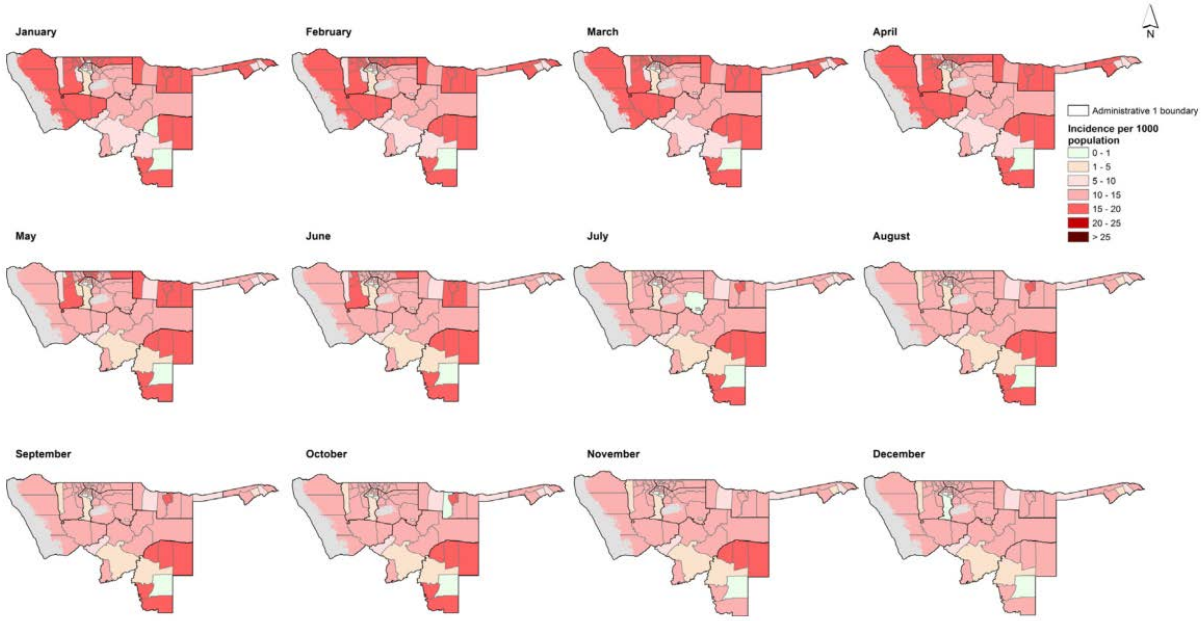


*Figure 4 a heatmap depicting the incidence of malaria in Nambia (****) at a county level by month in 2009. Adapted from ****

As discussed in the data section of this report many of the considerations discussed above are beyond my ability to access as an individual. I am approaching this project more as a proof of concept and application skillset than as a direct attempt directly (or accurately) predict Malaria incidence with publicly available data.

Next Steps

From here the plan is to apply machine learning algorithms (principally using SciKit Learn) to the available data with the goal of building a model for predicting the incidence of malaria given a set of factors. This would in theory then be able to be used to direct relief and medical efforts to the most stricken areas with more efficiency. The details of this will be discussed in the final report.

Appendix 1 : The Lifecycle of Malaria, Epidemiology, and related factors

Malaria: is a mosquito born disease (primarily the *Anopheles*) caused by parasites in the genus *plasmodium* (see below). These parasites are transferred to the human blood stream when the mosquito feeds. There are several subtypes of Malaria that are beyond the scope of this discussion, but the general life cycle can be seen below. **N.B. – An important factor in disease course i.e. difficulty in treating and severity of illness has to due with the species of malaria involved e.g. *P. falciparum* vs *P. vivax* etc. as the former causes far more deaths than the former. In this model I did not address either endemic species of malaria or mosquito.**

Figure 5 an Anopheles mosquito

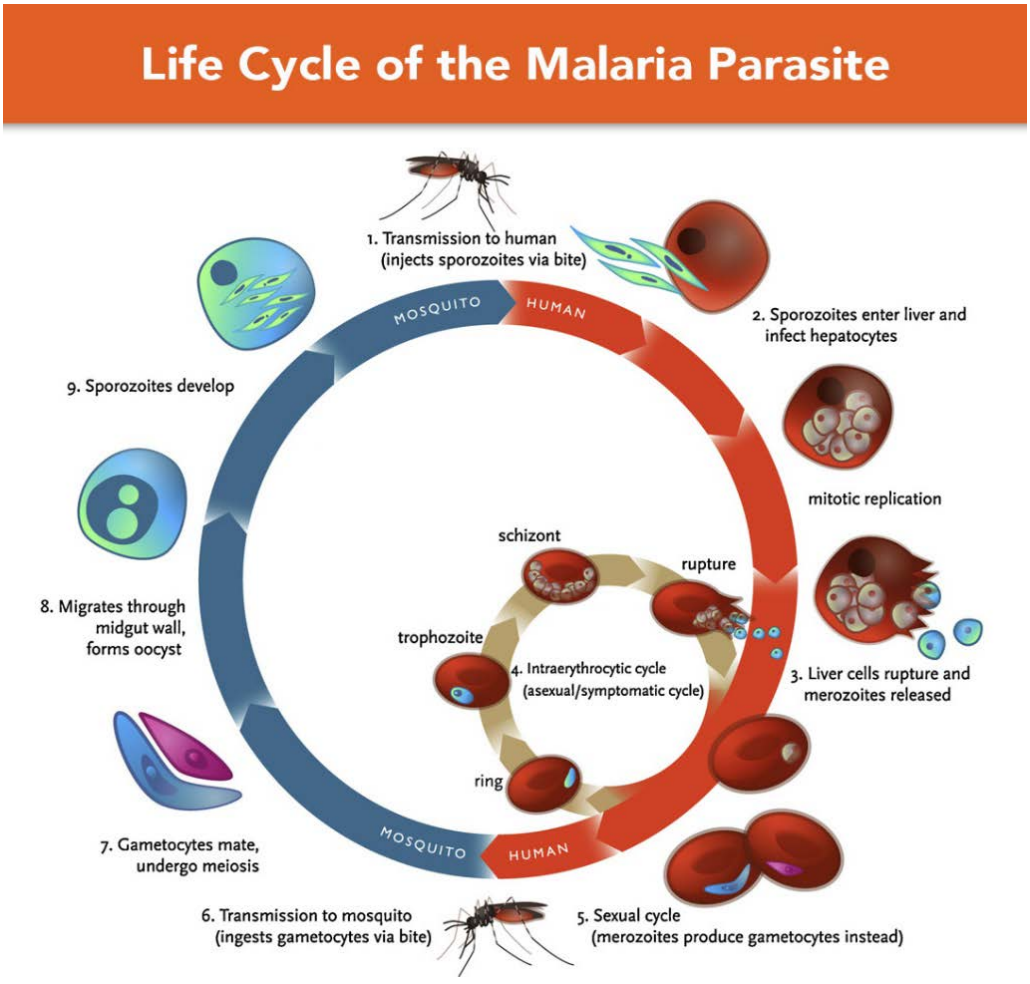


*Figure 6 Overview of the lifecycle (adapted from ***)*

Symptoms normally include fever, fatigue, vomiting, and headache. One classic feature of malaria is its cyclical fever (not always present). In the more pronounced cases it progresses to jaundice, coma, seizures, and death.

One of the main reasons tracking and/or prediction can be so useful is that there is approximately a two week inoculation time of the time of the bite to showing symptoms allowing for treatment intervention prior to symptoms even developing. Also, most forms of malaria can be treated relatively easily with modern medication (which does not cost much). Primary preventions is also a preferable route and this is why inset repellant nets are so often discussed as they cost < $2 a piece and have been shown to be high effective.

More information can be found here and here, interesting presentations here and here.

Associated factors

Certain factors have been consistently associated with the spread and impact of malaria: In general, these can be looked at from two categories: those that enable the spread of the vector i.e. mosquitos, and those that have a more classical association with the spread of disease e.g. human factors.

Weather and geographical: rainfall, temperature, humidity, and elevation. These factors will obviously vary from one part of a country to the next making the higher resolution variable discussed above all the more important. **N.B – with the changed being wrought by global warming the geographical regions classically associated with malaria are being shifted. We are also seeing some of this with the Zika virus.**

Human factors: Outdoor occupation, population density, urbanization (malaria incidence tends to be much lower in urban environments).

Other factors have been associated with a higher mortality rate i.e. more deaths. Some of these are distance from and access to proper medical treatment.

References

https://www.kaggle.com/teajay/the-fight-against-malaria

https://en.wikipedia.org/wiki/Malaria

https://www.unicef.org/media/media_20475.html

https://www.istockphoto.com/photo/anopheles-mosquito-gm153097831-21580214?esource=SEO_GIS_CDN_Redirect

https://cddep.org/tool/life_cycle_malaria_parasite/

https://malariajournal.biomedcentral.com/articles/10.1186/1475-2875-6-129

http://apps.who.int/iris/bitstream/10665/252038/1/9789241511711-eng.pdf?ua=1

https://www.unicef.org/media/media_20475.html

Alegana, V. A., Atkinson, P. M., Wright, J. A., Kamwi, R., Uusiku, P., Katokele, S., ... & Noor, A. M. (2013). Estimation of malaria incidence in northern Namibia in 2009 using Bayesian conditional-autoregressive spatial–temporal models. *Spatial and spatio-temporal epidemiology*, *7*, 25-36.

Noor, A. M., Kinyoki, D. K., Mundia, C. W., Kabaria, C. W., Mutua, J. W., Alegana, V. A., ... & Snow, R. W. (2014). The changing risk of Plasmodium falciparum malaria infection in Africa: 2000–10: a spatial and temporal analysis of transmission intensity. *The Lancet*, *383*(9930), 1739-1747.

Varun Kumar, Abha Mangal, Sanjeet Panesar, et al., "Forecasting Malaria Cases Using Climatic Factors in Delhi, India: A Time Series Analysis," Malaria Research and Treatment, vol. 2014, Article ID 482851, 6 pages, 2014. doi:10.1155/2014/482851