Annotated Bibliography

Sudheshna Bopati & Monét Norales

Source: https://www.cdc.gov/tb/publications/factsheets/statistics/tbtrends.htm
Author: N/A
Title: Trends in Tuberculosis
Venue: Center for Disease Control and Prevention
Year: 2021
Aim: Generic statistics of TB in the year 2020 in the USA. CDC provides statistics on TB that it is still endemic and encourages researchers to increase their efforts in detecting latent TB and strengthening current TB control practices to curb the infection
Conclusion: The article details the following statistics of the data

- US has seen a steep decrease in TB cases in the year 2020 reason predicted were probable underdiagnosis due to the prevalence of COVID or a true reduction
- TB is disproportionate amongst different factors such as demographic, health, and social factors with Non-Hispanic Asians at the highest 35.8% population in the US. Indicating TB is more common amongst people born in different countries other than the US and who share a lineage from different countries
- Pre-medical conditions can be a big factor in developing the infection. Diabetes mellitus at 22.5% was the most commonly reported medical risk factor for TB
- TB spread will also depend on people living in spaces where close contact with other humans is unavoidable. With 2.6% of 15 years or older residents in correctional facilities contracted TB

How does what they are saying inform how we design interventions/feedback for people?
This dataset/data is one of the earliest pieces of information that we found and played an important factor in making our decision to work toward TB. Also, the graphs in this article gave us an idea of how to present the epidemiological data on our webpage.

Source: https://www.who.int/news-room/fact-sheets/detail/tuberculosis
Author: Center for Disease Control and Prevention
Title: Trends in Tuberculosis, 2020
Venue: CDC Website
Year: 2021
Aim: key facts about TB and the infection rate statistics for the year 2020
Conclusion: The document gives us a gist of the key facts about the TB, some of them are presented here

- A total of 1.5 million died due to contracting TB in the year 2020((including 214 000 people with HIV).

- TB is caused by the bacteria mycobacterium tuberculosis which often affects the lungs
- 1/4th of the population has a latent TB that does not make them sick or infect others
- Older adults and HIV patients (18 times) more likely to develop TB infection
- Highest Tb occurs in the South-East Asian region with 43% of new cases.
- Multidrug-Resistant Tb is caused by bacteria that mostly do not respond to isoniazid and rifampicin. Second-line drugs can be used to cure MDR but treatment options are highly limited and 2 years of chemotherapy which is both expensive and toxic.

How does what they are saying inform how we design interventions/feedback for people?
The motivation behind working on TB is because in my team I'm from South-East Asia and know the impact that it has. And this sheet sums up our reasoning behind working on this specific bacteria.

Source: https://doi.org/10.1038/ng.2656
Author: Ford, C., Shah, R., Maeda, M. *et al.*
Title: *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis
Venue: *Nat Genet* 45, 784–790
Year: 2013
Aim: The authors are trying to understand why different strains are associated with different rates of multiple drug resistances. Compared strains from two lineages and also different strains within the same lineage and noted their difference in mutation rate.
Conclusion: The comparison of strains between two lineages of TB lineage2 (the East Asian) and lineage4 (Euro-American) have the following conclusions

- Lineage2 acquires drug resistance in-vitro more rapidly than lineage4, likely because of the genetic context of the TB strain that can impact the range of observed mutations conferring resistance to a single drug
- Also found significant genetic and phenotypic variations in the mutation rates of different strains within the same lineage
- The mutation rate per unit time is similar both in in-vitro and in-vivo. The reason behind this scenario is predicted as the mutation rate is dependent on time rather than the replication factor.
- Develop a predictive model to identify the TB mutation into multi-drug resistance before the onset of the treatment, concluding that the strains with higher mutation rates have an increased risk of drug resistance mutations.
- The model also predicts that bacterial burden is critical to determining the probability of drug resistance and early and active case detection is our best hope of curbing the multidrug resistance (MDR) epidemic

How does what they are saying inform how we design interventions/feedback for people?

This paper helps us understand that the mutation rate is relative to time and not a replication factor. And also lineage2 is mutating faster in-vitro compared to lineage4. This will be useful for analysis when trying to filter data. We would like to compare data for different strains not and for different continents/countries probably.

Source: https://www.who.int/teams/global-tuberculosis-programme/data
Author: Countries Reporting to WHO
Title: *Global Tuberculosis Report*
Venue: WHO website
Year: 2022
Aim: Gather data regarding TB from other countries to improve and expand the knowledge base.
Conclusion:

- First database: Downloaded the WHO TB estimates of TB mortality incidence burden that includes data disaggregation by age, sex, risk factors, HIV status, and rifampicin resistance.
- Second database: That gives us information about the different strains and different genes and their drug resistance data for the year 2020.

How does what they are saying inform how we design interventions/feedback for people?
This will help us with the dataset that we are looking for that has the strains data of the entire world, this could be a useful backend database that could be linked to the webpage.

Source: https://www.frontiersin.org/articles/10.3389/fmicb.2020.00667/full
Author: Kouchaki, S., Yang, Y., Lachapelle, et al.
Title: Multi-Label Random Forest Model for Tuberculosis Drug Resistance Classification and Mutation Ranking
Venue: Frontiers in Microbiology, Volume 11
Year: 2020
Aim: Tuberculosis strains can become resistant to multiple drugs called resistance co-occurrence. To predict the co-occurrence, a multi-label random forest (MLRF) ML model was compared with a single label random forest (SLRF) ML model for predicting phenotypic resistance and also identifying important mutations among 13402 TB isolates.
Conclusion: Resistance prediction and mutation ranking are important tasks in the analysis of Tuberculosis sequence data. Multidrug resistance (MDR) and First-line drug resistance (FDR) have been confirmed for some of the most frequent mutations capturing the association between the feature space and prediction of resistance and the credit goes to the MLRF. Although they suggest that limiting the mutation's data to the top 16-37 can make the algorithm run faster to predict the new mutations associated with MDR or FDR.
How does what they are saying inform how we design interventions/feedback for people?

The paper refers to the importance of looking at data of the TB strains with multiple drug resistance instead of one and they've ML concepts to prove that. Although we don't use ML in our dataset because of time constraints, this could be an amazing next-level approach for the database. And also it gives us the idea to filter out top multidrug resistance mutations for our database.

Source: *PeerJ*
Author: Kohl TA, Utpatel C, et al.
Title: MTBseq: a comprehensive pipeline for whole-genome sequence analysis of *Mycobacterium tuberculosis* complex isolates
Year: 2018
Aim: To develop a bioinformatics pipeline for the analysis of Mycobacterium tuberculosis complex (MTBC) isolates for a comprehensive antibiotic resistance profiling and outbreak surveillance
Conclusion: The researchers successfully developed an MTBseq pipeline that's fully automated and analyzes the Whole Genome Sequencing data. MTBseq provides a reference mapping-based workflow where a reference genome is mapped to all the reads and provides a list of variants, an SNP (single nucleotide polymorphism) distance matrix, and a FASTA alignment of SNP positions for use in phylogenomics, and identifies groups of related isolates. They also proved the function of the pipeline by comparing the results of phylogenetic analysis and cluster detection of 26 isolates with published findings and found similar results.
How does what they are saying inform how we design interventions/feedback for people? This doesn't help us with the current project since we are already considering a dataset where NGS sequence analysis has been performed but this could be a good introduction to understanding how the dataset that we are looking at is processed through NGS sequence analysis and the workflow within the pipeline to clean and convert data into a table format.

Source: https://www.sciencedirect.com/science/article/pii/S1471489217301571
Author: Koch, A., Cox, H., & Mizrahi, V.
Title: Drug-resistant tuberculosis: Challenges and opportunities for diagnosis and treatment
Venue: Current Opinion in Pharmacology, Volume 42, Page 7-15
Year: 2018
Aim: The paper aims to help people realize the gap between the poor diagnosis versus current drug treatment, especially for the multi-drug resistant strains of TB. Even though TB is known as a curable disease, the current statistics for TB with MDR is 56% which is low. Hence this paper addresses the key challenges in addressing the MDR issue and also explains the biology of TB drug resistance and disease pathogenesis.

Conclusion: The current resistance profile does not give a complete picture of the resistance and a more optimized version needs to be developed to diagnose resistance
The authors suggest that using standardized regimens such as considering a full resistance profile for all first-line and second-line TB drugs helps bridge the gap in poor treatments. And they also detect heterogeneity directly from biological specimens and therefore can identify any underlying drug resistance that could emerge during treatment.
How does what they are saying inform how we design interventions/feedback for people?
The paper has nothing that we need except that we could learn a lot of background about the genetics or biology of the drug resistance within TB

Source: https://www.nature.com/articles/s41598-020-65766-8
Author: Ghosh, A., N., S., & Saha, S.
Title: Survey of drug resistance-associated gene mutations in mycobacterium tuberculosis, Eskape, and other bacterial species.
Venue: Scientific Reports, Volume 10, Article 8957
Year: 2020
Aim: We are familiar with the idea that Treatment for TB includes broad-spectrum antibiotics. The authors developed a database called the Drug Resistance Associated Genes Database(DRAGdb) for specific drug resistance-associated genes across different antibiotics called ESKAPE(i.e. Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, and Enterobacter spp.) with a specific focus on TB in-addition to other bacteria. Considering the limitations in other databases they found a need to develop a new one that combines all organism-specific studies incorporated into a single database that can be used to analyze antibiotic resistance-associated mutations across different bacterial species.
Conclusion: The author's developed a database that can identify the common cause of resistance across different bacterial species and that helps us with identifying/developing novel drugs for targeting multiple bacterial infections at once.
How does what they are saying inform how we design interventions/feedback for people?
The papers have the similarity of developing a database that can give the drug resistance mutation data for also TB but the design of the webpage or the visualization of the data is not something we could go for. The database inputs a CSV file, we can probably use this data and visualize the data such that they might look at ours and use the idea to better visualize data.

Source: https://doi.org/10.1016/j.dib.2018.09.057
Author: Ismail, N., Omar, S. V., Ismail, N. A., & Peters, R.
Title: Collated data of mutation frequencies and associated genetic variants of bedaquiline, clofazimine, and linezolid resistance in Mycobacterium tuberculosis.

Venue: Data in Brief, Volume 20, Pages 1975-1983
Year: 2018
Aim: The goal of this publication was to catalog the mutations and, if possible, the mutation frequencies of the drug resistance of Mycobacterium tuberculosis to bedaquiline, clofazimine or linezolid from multiple other publications.
Conclusion:

- Compilation of mutations found for in vitro, in vivo and clinical drug-resistant strains
- Provides multiple tables for information on different strains, frequencies, concentrations, etc.

How does what they are saying inform how we design interventions/feedback for people?
This collection of data will be a great reference for mutation information on the drug resistant strains. Because the data were filtered for the type of mutation and the gene its located on, it will most likely be a base point for comparison to what information we get from other sources and for more detail as to the frequencies of these mutations.

Source: https://doi.org/10.1038/s41467-020-18699-9
Author: Colangeli, R., Gupta, A., Vinhas, S.A. *et al.*
Title: *Mycobacterium tuberculosis* progresses through two phases of latent infection in humans.
Venue: Nature Communications, Volume 11, Article 4870
Year: 2020
Aim: The aim of this paper was to report the rate trend of new mutations after two years of latent infection and further investigate the physiology of the latent infection. In this report, the researchers look at and analyze the possible mutation rates and generation time to see if there are any correlations with the latency period of the infection.
Conclusion: For the study of this paper, the researchers found that the mutation rates seemed to be very high during the early latency period and slow as it transitions to the late latency period. They also found that the mutation rate and one to two year latency might be correlated with secondary TB infections. Becasue of the difficult in observing the generation times, they were inferred by using the mutation rates and latency periods found during laboratory testing. Also, fro this study, the researchers noticed that the drugs only work on active bacteria as opposed to the dormant bacteria that contributes to the latency.
How does what they are saying inform how we design interventions/feedback for people?
This paper will be useful if there is a decision to create a visualiztion depicting the latency periods of different mutations or just the latency periods in relation to secondary infections. But if that is not done, this paper still provides decent insite to the influence of both phases of latent infections.

Source: https://academic.oup.com/jac/article/75/8/2031/5828363

Author: Kadura, S., King, N., Nakhoul, M., et al.
Title: Systematic review of mutations associated with resistance to the new and repurposed mycobacterium tuberculosis drugs bedaquiline, clofazimine, linezolid, Delamanid, and Pretomanid.
Venue: *Journal of Antimicrobial Chemotherapy*, Volume 75, Issue 8, Pages 2031–2043
Year: 2020
Aim: This aims to summarize and help to better understand the genetic aspects of the mutations related to drug resistance in the TB bacteria. The researchers were searching for links or associations between the genotypes and phenotypes involved in the mutations. They focused on experimenting with allele exchanges and also a few drug types.
Conclusion: From this study the researchers were able to identify mutations that are more frequent in response to specific drugs. Due to the drug resistance response, the researchers commented that keeping short term surveillance for genotype-phenotype links is needed when monitoring mutation occurrences and diversity in the mutation types.
How does what they are saying inform how we design interventions/feedback for people?
The association between drug and mutation that the researchers found/experimented with will affect the frequencies of the mutation occurrences based on the drug useage or prevalence. Because of this, it may be useful to try and create a visualization that can show the increase in frequency with the increase of a particular drug.