Big Data Analytics in Bioinformatics

Spurthi G S

CSE, BNM Institute of Technology, Bengaluru, Karnataka, India

***Abstract*— Tuberculosis is the ancient and global disease, which is found worldwide. TB is the infectious bacterial disease which affect both humans and animals due to growth of nodules in the tissues (mainly Lungs).Big data analytics in healthcare is evolving into a promising field for providing insight from very large data sets and improving outcomes while reducing costs..The paper highlights the potential of bigdata to identify contiguous infection and to hypothesize a Data application engine that not only provides mapping between the spread of animal-human tuberculosis but some of the neglected reason about the spread of the disease.**

***Keywords*— Big data; Tuberculosis; Mycobacterium tuberculosis;**

I. INTRODUCTION: ANIMAL AND HUMAN TUBERCULOSIS AND BIG DATA

Tuberculosis is one of the oldest diseases that have coevolved along with human from many thousands of years. Robert Koch was awarded with Nobel prize for discovering Mycobacterium tuberculosis which is responsible for causing tuberculosis in 1905.Tuberculosis in human is caused by Mycobacterium Tuberculosis .Other members that belong to this family include M.bovis which is the main cause for tuberculosis in animals. M.microti and M.africanum are very rare bacillus which cause the infections. Human may become infected by M.bovis usually due to intake of milk, milk products or meat from the infected animal.6% of death of humans in preantibiotic era was due intake of infected products because there was no concept of pasteurization. Even though bacilli was identified nearly 130 years ago, a definitive understanding of pathogenesis of this disease is still deficient .Infection with TB can result in two stages: asymptomatic latent tuberculosis infection (LTBI) or tuberculosis disease. If left untreated, the mortality rate with this disease is over 50%. The Case facility ratio(CFR) in 2015 was carried out which shows most of WHO African region are high suffers. This shows considerable inequalities among countries in access to TB diagnosis and treatment that need to be addressed. So if everyone with TB had a timely diagnosis and high quality treatment the CFR would be low in all Countries.

In order to achieve Low CFR Bigdata analytics can be used. Bigdata analytics in healthcare is driven by mandatory requirement and have the potential to improve the quality of healthcare delivery meanwhile reducing the cost which is one of the important factor to fasten the process of TB detection and prevention. Bigdata hold the promise of supporting a wide range of medical and healthcare functions to derive previously untapped intelligence and to get insight from data sets to address many new and important challenges that will be keep on arising day by day. Consuming un-pasteurized milk act as a mode of transmission of "Bovine Tuberculosis". In a normal situation, around 10% of infected individuals will suffer from this disease once in their life time. If people with the disease fail to take treatment, 50% of PTB+ may die within 5 years, whereas 25% may remain chronically ill with infectious "Tuberculosis" and the rest 25% may get cured automatically. An approximate of 2 million people die every year due to "Tuberculosis". Of the 22 "Tuberculosis" high burden countries responsible for 80% of total universal "Tuberculosis" saddle, 9 are in Africa whereas among the 15 countries with the highest TB frequency rate per capita, 12 are in Africa[3]. Regarding awareness of study participants on "Bovine Tuberculosis, 82% of people are unaware that "Tuberculosis" is transmitted by livestock, or through some of the products, whereas only 18% had heard about it. Because of close contact of humans with livestock, humans could be at risk of acquiring "Bovine Tuberculosis". As per the study, vast majority (91%) of participants reported to consume raw milk. Bovine Tuberculosis is basically started in areas where public are in close get in touch with livestock(Abdi A Gele,2009). This is supported by the reality that risk elements for "Bovine Tuberculosis" spread such as using up of raw milk and distributing same variation (same fence) with livestock were found highly universally among this cluster. (Houser, 2015) (Abdi A Gele,2009)

II. GLOBAL SCENARIO

Tuberculosis is one of top 10 deadliest contagious disease. According to Estimate made by WHO in 2015. 1.4 million people died because of tuberculosis and additionally 0.4 million deaths resulting from TB disease among HIV positive. Even though TB is slowly declining each year according to estimate 37 million lives were saved between 2000 and 2013 through effective diagnosis and treatment. But most shocking factor was reported by WHO in its global tuberculosis report 2016 which provides an assessment of TB epidemic and progress in TB diagnosis i.e., the epidemic spread is larger than estimated count (especially in India)

In Southeast Asian(SEA) region of WHO is home to 25% of the world human population and the miserable fact is that with 30% of the world's poor belong to this region. Due to lack of better health infrastructure SEAR suffers from high risk of

both communicable and non-communicable diseases. Progress in global health will not be possible without visible progress in this region. Six of the 14 million deaths in this region are caused by communicable diseases like TB is one out of them. Even though the various strains of TB are quite difficult to catch because it usually requires prolonged exposure to the infected person or animal. Majority of person(about 9 out of 10) whoever infected with the bacteria causing TB may or may not develop a symptoms or disease .Around 10% of infected may develop a disease. Bacteria that causes TB can survive in the human body in inactive state for decades and develop a disease some time later when the immune system become weak may be during old age or due to poor living conditions. The TB disease can be cured but it requires intake of long course of antibiotics. WHO as came up with the End TB strategy ,the overall goal is to "end the global TB epidemic".

The End TB Strategy has 3 higher level, overarching indicators and related targets and milestones. The 3 indicators are, The number of TB deaths per year ,The Tb incidence rate per year, The percentage of TB affected households that experience catastrophic costs as a result of TB diseases .The trajectories of TB incidence and TB deaths are required to read End Tb strategy milestones and targets shown in figure 1.

The first milestones of the end TB strategy set for 2020 are a 35% reduction in the absolute number of TB deaths and 20%reduction in the TB incidence rate. Compared with levels in 2015.To reach these milestone, The TB incidence rate needs to be falling by 4-5% per year globally by 2020 and the proportion of people with TB who die from the disease needs to be reduced to 10% globally by 2020 .Worldwide in 2015 there were an estimated 10.4 million incident TB cases.6countries accounted for 60% of global total India, Indonesia, China, Nigeria, Pakistan & South Africa

VISION	A WORLD FREE OF TB — zero deaths, disease and suffering due to TB			
GOAL	END THE GLOBAL TB EPIDEMIC			
INDICATORS	MILESTONES		TARGETS	
	2020	2025	SDG 2030*	END TB 2035
Percentage reduction in the absolute number of TB deaths (compared with 2015 baseline)	35%	75%	90%	95%
Percentage reduction in the TB incidence rate (compared with 2015 baseline)	20%	50%	80%	90% (approximately 10 per 100 000 population)
Percentage of TB-affected households experiencing catastrophic costs due to TB (level in 2015 unknown)	0%	0%	0%	0%

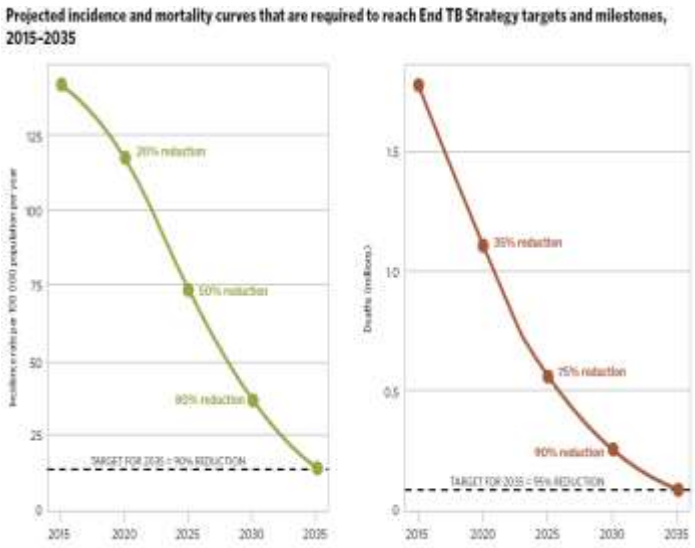

Figure 1: End Tb strategy milestones and targets

III. INDIAN SCENARIO

In India, TB has been mentioned in the *Vedas* and the old *Ayurvedic* scriptures. Historically speaking, fight against TB in India can be broadly classified into three periods: early period, before the discoveries of x-ray and chemotherapy; post-independence period, during which nationwide TB control programs were initiated and implemented; and the current period, during which the ongoing WHO-assisted TB control program is in place

IV. MOTIVATION

Big data has been defined as the collection of complex data sets difficult to manage and process using traditional applications. Big data is gaining interest day by day because, First, the datasets is growing by the sequence of magnitude which holds valuable insights. Traditional data warehousing model will not hold good in scaling and analyzing large volumes of datasets. Second, the analyzing the large amount of data requires parallel processing. Traditional data warehousing model lacks in achieving parallel processing and analysis of large amount of data. Third, there are many varieties of datasets some of which are also referred to as inconsistencies with the datasets. Managing such data effectually is hard with the traditional data warehousing model. Fourth, with the traditional data warehousing model taking care of large volumes of datasets is a knotty process and requires a good quality from the datasets. Hence, the study of the "Big Data" technologies has introduced an application engine capable of storing, managing & analyzing the "Tuberculosis spread Between Humans & Animals" (TBHA) data in order to provide a confirmative scaling between the spread of Tuberculosis between animals & humans.

Many researchers concur that big data which is in its early days has most of its potential for value Creation still unclaimed. Big data rely on large data sets and predictive analysis to produce insight for decision making in the healthcare sector. The power of big data to provide insights to problems in many disciplines has seen major investments and excitement in big data. Whilst many sectors have embraced big data, the health sector is lagging behind on the use of big data. The potential value creation of big data is in aggregating individual data sets into algorithm to provide insights rare with individual data sets.

Most researchers concur that big data is the cornerstone of modern epidemiology because of the availability of computational and analytical tools to deal with complex large data sets. A study by Harvard revealed that big data have a potential of realizing $300 billion annual savings in the US health sector. In addition to the associated cost savings, in some cases it is a matter of life and death without big data. The combining of data sets allows an opportunity to learn the relationship between different TB risk factors. Many researchers concur that big data have potential to generate a hypothesis which helps to gain insights on TB. The lack of data on the impact of various risks (silicosis, HIV infection, malnutrition, diabetes, smoking, crowded living and indoor pollution) poses a challenge to the understanding of TB epidemic. The availability of information helps to narrow the focus on better preventive and control interventions and focus more on high prevalence low risk factors than low prevalence high risk factors (prevention paradox). The big data therefore have the potential to enable quicker interventions to hot spots through data driven monitoring of TB. The ability to develop complex big data analytical techniques will enable to understand the cause and effect Tuberculosis. Some of big data capabilities include reporting (What happened?), monitoring (What is happening now?), data mining (Why did it happen?), evaluation (Why did it happen), predictive (What will happen?) [5]

The traditional "data" warehousing model are relational databases it will act as data analysis tool. The data warehousing model is facing a key challenge such as scaling large volumes of data and analyzing data with high velocities and variety of unstructured formats.

V. OBJECTIVE AND PROBLEM

Objective of this case was to understand a "Big Data" application engine that not only provides a confirmative mapping between the spread of Tuberculosis between animals & humans but also tells us about some of the neglected reason of the spread of the disease.

Tuberculosis is a major concern across the world not only in humans but also in the livestock industry, zoo animals and wildlife. The latest Big Data Analytic techniques are used to gain valuable insights on how Tuberculosis in animals infects humans and vice versa by analyzing the TBHA i.e. humans/animals infected by Tuberculosis due to others' data sets.

VI. BIGDATA AND TB

Zettabytes of unprocessed information on Tuberculosis could provide ideas for everything ranging from preventing the disease to reducing the analytical costs. The incoming data that grows up in multiplicity, volume, velocity and variety holds the potential of valuable insights (Sagiroglu, 2013).

Some of the characteristics of "Tuberculosis spread Between Humans &Animals" (TBHA) data that is the input to the study is presented in figure 2.

Hence any data which a traditional data warehouse cannot manage is called Big Data. Big Data is dormant in making improvements in operations, make faster and more intelligent decisions. New techniques, tools and architectures are needed for Big Data in order to solve new problems in a better way and also to solve old problems as the database management tools and other data processing tools are not much sufficient in process and the analysis of the data[6]

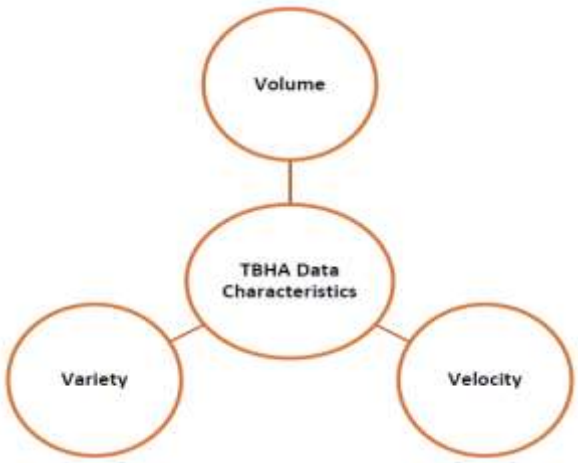

Figure 2:Spread of TB

VII. EPIDEMIOLOGY APPROACHES TO TUBERCULOSIS

According to Roy and Chauhan [6] there are three epidemiology approaches to tuberculosis which are: analytic epidemiology (dealing with risks factors associated with the agent), descriptive approach (dealing with prevalence and incidence of tuberculosis infection) and lastly the predictive approach (dealing with the forecasting of the tubercular epidemic). They state that the epidemiology is important for the implementation of national tuberculosis control programmes. They add that in order to understand the epidemiological basis of tuberculosis control, there is a need to understand the dynamics of the disease and major determinants of epidemiology of tuberculosis. They proposed

a four step model (exposure, infection, disease and death) as in figure for understanding epidemiology tuberculosis control [7]

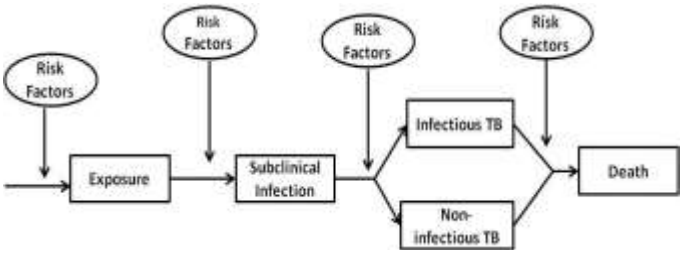

Figure 3:Four step model

According to the model Figure 3 , an exposure to an infection case is necessary to acquire the tuberculosis infection. The understanding of risk factors that lead to infection is dealt with by analytical epidemiology. The distribution and frequency of the disease in a given community is dealt with by the descriptive approach. Finally the forecasting and modelling of the epidemic based on observation from the past is dealt with by the predictive epidemiology. The insight from understanding risk factors of a given community is important to develop effective tools to prevent and control tuberculosis. They state that predictive and descriptive epidemiology is essential for efficient and effective tuberculosis control programmes.

According to Khoury and Wagener [7] the prevention of common diseases relies on identifying risk factors and implementing intervention in high-risk groups. They add that the classical epidemiological paradigm of searching for "risk factors" and intervening in high-risk groups has enjoyed much success in controlling and preventing many infectious diseases. Boonstra and Broekhuis [8] emphasize that the implementation of an integrated health information systems has the potential for improving the quality and efficiency of health delivery system across the world. Paul et al. [9] state that integrated health information systems have the potential to offer economic benefits through efficiency savings via appropriate data management to identify potential bottlenecks in the provision and administration of care, which can become more predictable.

VIII. SPREAD OF BOVINE TUBERCULOSIS THROUGH ANIMALS

"Bovine Tuberculosis" spreads from sheep to Pigs. Pigs are uncovered to bovis virus and bovis was a common reason of "Tuberculosis" in pigs. Disease counts in pigs typically reflect those in cattle, and occurrence in the order of 20% has been recorded in some pig populations. The oral route is the most important route of virus in familial pigs, most of them are frequently infected by feeding milk, milk products or from infected cows.

"Bovine Tuberculosis" spreads from Cattle to Goats. In Australia, goats are rarely set up with cattle. In this goats were grazing with unhealthy cattle (with a prevalence of 35%) at different high stocking rates. Although the goats appeared scrawny, disease was only recognized in a single animal after skin trying and repeated culture of reactor lymph nodes.

Bovine TB in livestock. Mycobacterium bovis infection in horses is not common; when disease occurs, the primary lesions are originate in the abdomen, suggesting intake as the cause of infection. The occurrence of TB in livestock is very squat in countries with a national programme to expunge tuberculosis. Information from new trial contagion suggests the horse is fairly loath to infection with M. Bovis[20].

Dogs are infected by Bovine TB. Although new studies have disclosed that dogs are evenly abnormal to M.bovis and M. tuberculosis infection summary of bacteriological studies have indicated that tuberculosis in dogs is more normally caused by TB

IX. CEASING ANIMAL HUMAN TB OUTBREAK BY BIGDATA

Big data has really commutated epidemiology. Even though there are some important challenges to survive with, there are many behavior to make use of data in a new pioneering way that could result in successfully defy the animal human TB. Big data can be used to get the disruption under power and to evade animals and humans from getting unhygienic.

One of the most vital responsibilities in stopping an tuberculosis disruption in both animal and human is product all the people who were in contact with an infected person and glance them for the period of the incubation era[15].

Big data actually consists of collecting huge amounts of information from a large range of sources. Once the information is organized and collected, big data analytics takes over, mining that data and searching preceding unseen patterns and tendency that would take humans years to discover, if they could do it at all. "Big data analytics" also removes any data that is allowed unimportant or irrelevant, further helping in establishing the data that matters

X. CONCLUSION

The only way big data can be impressive in fighting tuberculosis in both animal and human is by collecting all the necessary data from as many number of sources as possible, or what's known as multi-center ingest. This technique and the technology to conclude it were simply not possible even a few years ago. Now, researchers and data experts can gather information from social media, hospital records, flight records, and even media reports. From there, they can more exactly predict where the disease may spread, which is a basic step in stopping an ongoing outbreak.

One of the key neglected reasons for "Tuberculosis" outbreak deducted from the results section are as follows: Identification of Tuberculosis in animals, Not taking into consideration the increase of Tuberculosis between mammals and humans, particularly within rural areas. Limited availability of data & study on "Tuberculosis spread between animals and humans", Use of traditional error-prone methods of data collection, storage and management leading to inaccurate, incomplete and wrong data, Drug-resistant Tuberculosis development in animals and its effect on humans, Need for treatment solutions for addressing Tuberculosis in animals.

Hence, when the TBHA data collected, it is concluded that, the treatment for Tuberculosis in general should be improved across the globe since the total count of "Tuberculosis spread among animals and humans" across different countries is increasing between 1990 and 2010. By comparing Africa and India, we can say that the treatment for "Tuberculosis" received in India for animals and humans is better than that received in Africa

REFERENCES

[1] Abdi A Gele. "Pastoralism and delay in diagnosis of TB in Ethiopia", BMC Public Health, 2009 Ahmed El Idrissi (FAO), Elizabeth Parker (FAO) "EMPRES Transboundary Animal Diseases Bulletin".

[2] Ayele, W. Y., S. D. Neill, J. Zinsstag, M. G. Weiss, and I. Pavlik. "Bovine tuberculosis: an old disease but a new threat to Africa." The International Journal of Tuberculosis and Lung Disease 8, no. 8 (2004): 924-937.

[3] Beyond the hype: Big data concepts, methods, and analyticsAmir Gandomi∗, Murtaza Haider. Cosivi, O., J. M. Grange, C. J. Daborn, M. C. Raviglione, T. Fujikura, D. Cousins, R. A. Robinson, H. F. Huchzermeyer, I. De Kantor, and F. X. Meslin. "Zoonotic tuberculosis due to Mycobacterium bovis in developing countries." Emerging infectious diseases 4, no. 1 (1998): 59.

[4] Etter, Eric, Pilar Donado, Ferran Jori, Alexandre Caron, Flavie Goutard, and Francois Roger. "Risk Analysis and Bovine Tuberculosis, a Re-emerging Zoonosis." Annals of the New York Academy of Sciences 1081, no. 1 (2006): 61-73.

[5] Fournier, A., I. Young, A. Rajić, J. Greig, and J. LeJeune. "Social and Economic Aspects of the Transmission of Pathogenic Bacteria between Wildlife and Food Animals: A Thematic Analysis of Published Research Knowledge." Zoonoses and Public Health (2015).

[6] Gortazar, Christian, IratxeDiez-Delgado, Jose Angel Barasona, Joaquin Vicente, Jose De La Fuente, and Mariana Boadella. "The wild side of disease control at the wildlife-livestock-human interface: a review." Frontiers in Veterinary Science 1 (2015): 27

[7] Houser, Christine M. "General Infectious Disease Question and Answer Items." In Pediatric Infectious Disease, pp. 27-187. Springer New York, 2015.

[8] Humblet, Marie-France, Maria Laura Boschiroli, and Claude Saegerman. "Classification of worldwide bovine tuberculosis risk factors in cattle: a stratified approach." Veterinary research 40, no. 5 (2009): 1-24.

[9] Janert, Philipp K. Data analysis with open source tools. "O'Reilly Media, Inc.", 2010.

[10] Japhet, I. S. H. A. K. U., W. U. N. A. M. I. R. Jalyson, FrouzanNaghshbandi,Bahman Zarrinjooee,ArezuNamadi,RoohollahBahmaniZargari, Mansur AminiLari et al. "The Presence Of Mycobacterium Species In Raw Milk Obtained From Lactating Cows In Lewa And Duda Of Vimtim Wards, Mubi North." Prevalence 2015 (2015): 4-1.

[11] Michalak, Kathleen, Connie Austin, Sandy Diesel, M. J. Bacon, Phil Zimmerman, and Joel N. Maslow. "Mycobacterium tuberculosis infection as a zoonotic disease: transmission between humans and elephants." Emerging infectious diseases 4, no. 2 (1998): 283.

[12] Michel, Anita Luise, Borna Müller, and Paul David Van Helden. "Mycobacterium bovis at the animal–human interface: A problem, or not?" Veterinary microbiology 140, no. 3 (2010): 371-381.

[13] Oh, Peter, Reuben Granich, Jim Scott, Ben Sun, Michael Joseph, Cynthia Stringfield, Susan Thisdell et al. "Human exposure following Mycobacterium tuberculosis infection of multiple animal species in a metropolitan zoo." Emerging infectious diseases 8, no. 11 (2002): 1290-3.

[14] Olston, Christopher, Benjamin Reed, Utkarsh Srivastava, Ravi Kumar, and Andrew Tomkins. "Pig latin: a not-so-foreign language for data processing." InProceedings of the 2008 ACM SIGMOD international conference on Management of data, pp. 1099-1110. ACM, 2008.

[15] Sagiroglu, Seref, and DuyguSinanc. "Big data: A review." In Collaboration Technologies and Systems (CTS), 2013 International Conference on, pp. 42-47. IEEE, 2013.

[16] Thakur, Aneesh, Mandeep Sharma, Vipin Katoch, Prasenjit Dhar and R Katoch "Study on the prevalence of Bovine Tuberculosis in farmed dairy cattle in Himachal Pradesh", Veterinary world 2010.

[17] Thoen, C.O., LoBue, P., Enarson, D.A., Kaneene, J.B. & de Kantor, I.N. 2009. Tuberculosis: a re-emerging disease of animals and humans. Veterinaria Italiana, 45(1):135–181.

[18] Thusoo, Ashish, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff, and Raghotham Murthy. "Hive: a warehousing solution over a map-reduce framework." Proceedings of the VLDB Endowment 2, no. 2 (2009): 1626-1629.

[19] Van Embden. "Use of various genetic markers in differentiation of Mycobacterium bovis strains from animals and humans and for studying epidemiology of bovine tuberculosis." Journal of clinical microbiology 32, no. 10 (1994): 2425-2433.

[20] Van Soolingen, Dick, P. E. De Haas, Jan Haagsma, Tony Eger, P. W. Hermans, V. Ritacco, A. Alito, and J. D.