Wireless Personal Communications (2021) 117:87–108
https://doi.org/10.1007/s11277-020-07040-8

An Optimized Integrated Framework of Big Data Analytics Managing Security and Privacy in Healthcare Data

Ritu Chauhan[1] · **Harleen Kaur**[2] · **Victor Chang**[3]

Published online: 19 February 2020

Abstract
Big data analytics has anonymously changed the overall global scenario to discover knowledge trends for future decision making. In general, potential area of big data application tends to be healthcare, where the global burden is to improve patient diagnostic system and providing patterns to assure the privacy of the end users. However, data constraints exists on real data which needs to be accessed while preserving the security of patients for further diagnostic analysis. This advancement in big data needs to addressed where the patient right needs to maintained while the disclosure of knowledge discovery for future needs are also addressed. To, embark and acknowledge the big data environment its adherently important to determine the cutting-edge research which can benefit end users and healthcare practioners to discover overall prognosis and diagnosis of disease while maintaining the concerns for privacy and security of patient data. In current state of art, we tried to address the big data analytics approach while maintain privacy of healthcare databases for future knowledge discovery. The current objective was to design and develop a novel framework which can integrate the big data with privacy and security concerns and determine knowledgably patterns for future decision making. In the current study we have utilized big data analytical technique for patients suffering from Human Immunodeficiency Virus (HIV) and Tuberculosis (TB) coinfection to develop trends and detect patterns with socio economic factors. Further, a novel framework was implemented using unsupervised learning technique in STATA and MATLAB 7.1 to develop patterns for knowledge discovery process while maintain the privacy and security of data. The study overall can benefit end users to predict future prognosis of disease and combinatorial effects to determining varied policies which can assist patients with needs.

Keywords Big data · Big data analytics · Security and privacy · Healthcare databases

✉ Harleen Kaur
harleen.unu@gmail.com

[1] Amity University, Noida, India

[2] Department of Computer Science and Engineering, School of Engineering Sciences, and Technology, Jamia Hamdard, New Delhi, India

[3] School of Computing, Engineering and Digital Technologies, Teesside University, Office P1.07a, Phoenix Building, Stephenson Street, Middlesbrough TS1 3BA, UK

1 Introduction

The expansion of digital knowledge is engendering vast amount of data in exabytes, hence generating challenges among the researchers and scientists to design and develop automated technology. However, big databases are creating painstaking efforts among researchers to handle and develop knowledge-based technology for future decision making. In, fact the EHR (electronic Health Records) has undergone dramatic shift which has perpetually changed the global scenario of healthcare application domain. The shift has occurred due to novel ICT (Information and Communication technology) which has changed the era of traditional tool to new generation of sensor-based technology, imaging, scanning and other technological advancements. The focus is to retrieve hidden patterns and information from big data bases using new intervene technology which can evidently improve the clinical decision making while maintaining privacy and security of patient's data. So, big data analytics tends to be anticipated technology which can be widely applied in healthcare application domain in varied areas which include insurance fraud detection, treatment, predictions of disease and identifying factors related to healthcare costs [1–5]. Hence the ultimate goal is to deliver an effective and efficient treatments to benefit end users for future decision making.

Big data analytics can be foreseen as new adoptive IT based approach which can render wide benefits to healthcare practioners for transformation in better clinical decision making. Indeed, the fear exists when the privacy and security of big data is a concern for healthcare. Thus, challenge is to handle complex and voluminous nature of data, with the next level of atrocities is to establish and protect the patient level data. Such series of complex tasks are established with new novel algorithmic technology and standards [5–7].

Big data analytics with healthcare technology can vastly improve the efficiency and effectiveness of patient care by providing the insights of data with utmost security and privacy. This can facilitate the various process healthcare which include patient data flow, overall patient stays during hospitalization, insurance data and other cost related features to improve the quality of care. In general, big data analytics is an imperative technology to generate impounded outcomes to reduce the global burden of disease with focus on data privacy and security. Hence, data analytics techniques bound to have extensive vision where the paradigm is to generate and develop secure and sustainable tool for varied application domains which include healthcare, GIS (Geographical Information system), imaging, industry, banking, and others [8–11]. In technological context there exist huge advancement in electronic databases with volume and complexity, herewith the knowledge discovery is discovered as an exponential tool to analyse data. However, knowledge discovery is explicit term which is radical in nature and prone to detect hidden patterns and knowledge which can benefit end users for varied application domains with privacy and security concerns.

The big data analytics encompassing effect tends to discover hidden patterns which has been classified using various algorithm technique such as Artificial Neural Networks where the technique usually works on the concept of biological model where the brain neuron send the signals to different parts for appropriate actions in similar context the predictive model design the data to gather the information from training data sets and depicts the overall chances of prognosis of disease for future outcomes, to classify the data Decision trees are variably utilized to discover a tree like structure where the tree is depicted as per training datasets. The Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) are two best known

techniques to create decision tree. The K-Nearest neighbour method identify the closest neighbour with respect to distance matrix and classify every record with combination among the groups of the k records. A new era big data technique has evolved with genetic algorithms where the concept is to evolve new modified algorithms to substantiate the genetic combination for natural selection and mutation of evolution for discovery of hidden patterns with utmost security and privacy [12–16].

The concern big data is facing is security and privacy in the domain of healthcare where the existence of advancement in emergence threats which can occur due to gaps and disclosures in adaptive information systems. Although, this has created an immense challenge among the researchers and scientists around to globe to deal with complimentary issues of security and privacy. As, we know healthcare is very sensitive issue where the patient data records should be kept at most private regardless of any changes in policy and guidelines [17–21]. Thus, patients are suffering from varied critical disease which include HIV, AIDS, Cancer, Tuberculosis and others where the patients do not want to share details with any organization which can hamper the basic socio needs and well-being of patient [8, 22–26].

This invasion in privacy and security of data should be assessed as a persistent threat where it should be discussed as a critical issue in context with patient privacy. This threat has raised several questions and suggestive studies where the privacy and security were the foremost challenge among the researchers and scientists. Hence, privacy is the major concern in healthcare databases which must be dealt on highest priority for decision making. As a result, developers are complimentary their studies with big data analytics to assess the nature of data and provide comprehensive technology for analysis.

In general, big data analytics provides variable opportunities which can benefit healthcare practioners to transform the application to retrieve effective and efficient patterns, but it has multifaced challenges which include privacy and security, complexity, high dimensionality and other. Hence, privacy and security are embarked issues which raises serious concern to protect patients records from cyber threats. The concerned are elevated among healthcare practioners to determine an appropriate technology which can adequately safeguard the privacy and security of organization. Indeed, the current gaps in technological implication has motivated to regulate the breach in security and effective usage of healthcare data. In the current study of approach, we discuss the work related to big healthcare databases and presents jeopardies which can adhere with security and privacy of data.

Thereafter, we tried to address the big data analytics approach while maintain privacy of healthcare databases for future knowledge discovery. The current objective was to design and develop a novel framework which can integrate the big data with privacy and security concerns and determine knowledgably patterns for future decision making. The novel framework is implemented using unsupervised learning technique in STATA and MATLAB 7.1 to develop patterns for knowledge discovery process. In the current research we have utilized Big data Analytical techniques in healthcare databases with patients suffering from HIV (Human Immunodeficiency Virus) and TB (Tuberculosis) coinfection to develop trends and detect patterns with socio economic factors to deal with privacy and security risks in healthcare databases. The focus of study was to retrieve hidden patterns and information from big data bases using new intervene technology which can evidently improve the clinical decision making while maintaining privacy and security of patient's data. Thereafter, big data analytics tends to be anticipated technology which can be widely applied in healthcare application domain in varied areas which include insurance fraud detection, treatment, predictions of disease and identifying factors related to healthcare costs.

The overall paper is discussed as Sect. 2 will overview the impact of big data on healthcare with concerns on privacy and security with focus on past literature study. A novel framework is designed and developed to effectively and efficiently capture big data from various resources in context to maintain the privacy and security of data and detect hidden patterns for clinical decision making for retrieval of information from large scale databases in Sect. 3. However, effective and efficient implementation of corresponding framework is discussed in Sect. 4. In last Sect. 5 conclusion is deliberated.

2 Privacy as Concern in Big Healthcare Databases

In the high-end epoch of technology, seamless available resources have created an unprecedented growth of databases which is pronounced with a new term as "Big Data". The data stored in varied resources consists of unstructured, structured, text, images, such complexity of data is difficult to be managed with traditional statistical technology. Hence, analytical technique needs to be comprehended with upcoming challenges of big data in varied application techniques.

In general, big data is vastly applied in diverse area of healthcare, which has widely enabled healthcare practioners and researchers to gain insight of data for better decision making. But the challenge exists to enable a better efficiency patient model which can benefit overall cost and privacy of the patient. A similar approach conducted in the project "WORLDII" an initiative conducted by New Zealand Privacy Commissioner aims to legalize the privacy of data flow from one consecutive location to another [27]. The patient's data is a sensitive issue where the prioritization is on big data with privacy needs. A data directive is issued by European council for data flow in compliance with e-health to assure that personal data is processed with utmost security and credential detail or private data should not flow at free access.

Foremost, to enable a platform which should be followed among the organization with privacy of data a guideline is issued by Organization of the Economic Cooperation and Development (OECD). The policy assures that individual patient data or personal data (PD) should be secured, so patient should not suffer any arbitrary losses [28]. The policies are minimum standards laydown in eHealth sector and should be followed for protection and privacy of patient data.

In general, the security and privacy among the big data with healthcare has limitless research where getting access and control of data is a complex situation. In this regard, several organizations in healthcare must have security measures so they can protect the data flow which can be embedded for integrated hardware and software system among big data. The data lifecycle in the security is a new contest of security in big data where it is as refereed in three aspects which include data security, the control and access of the data with relevant information security [9]. The data lifecycle was established to enable effective and efficient decision making in context with data.

Moreover, the development technological need is to assure to determine the effective and efficient patterns from big healthcare data in concern with maintaining the privacy and security of data. Hence, big data analytics is widely anticipated technology where the potential is to retrieve hidden patterns and information from raw datasets. The patterns detected can be exploited for real world application domain where the decision support mechanism can benefit healthcare practioners. As we know the medical databases has changed vastly from the past two decades due to intervene IT based

technology where new type of datasets have evolved like the EHR (Electronic Health Records), imaging, Radiation data are generated at an amicable speed which is difficult to handle via traditional tools. Hence, big data analytics has ability to handle the complex nature of data and detect knowledgeable patterns for decision making [5–7, 12].

While the healthcare sector, is transmitting the data at utmost speed and lacks the delivery support system to generate the validated predictive results in concern with privacy and security of big data. In fact, it's complicated matter where the healthcare practioners are unaware of the threats which are induced and can susceptibly hamper the patient personal data. Whereas, implementing the big data with security concerns is the trivial area of research among the scientists and healthcare practioners [29–34].

In past, several data mining techniques are involved to breach the security and privacy of data to gather the sensitive data and publicly revealed the secure or personal details. So, security remains a complex task where sophisticated technology should be adoptive to analyse the big data. In current study of research big data analytical approach is applied with privacy among the healthcare data to provide minimal access to personal data of patient and retrieval of effective and efficient patterns for knowledge discovery [21, 35–39].

In the current approach we have applied the big data analytic techniques for patients suffering from Tuberculosis (TB) and HIV while keeping in concern with maintaining privacy of patient personal data. As we know TB affects people of all ages around the globe. But in year 2012, about 80% TB cases were reported from just 22 countries, showing a greater prevalence of the disease in some countries as compared with other countries. This increased prevalence can be due to multiple reasons including population, lifestyle of the people, major occupation, gene pool of the region and other socioeconomic factors. HIV and Tuberculosis have always been linked to the economic and financial status of the individuals [15].

Thus, several data analytics studies are interlinked with each other to determine the cause of TB and its prognosis with HIV [1–4], but privacy remains an exclusive cause of study. However, in healthcare several procedures are discussed to maintain the privacy of data which includes:

1. Authentication

The authentication of data and user with app authentic behaviour is a complex situation where every organization needs to embark and determine the confirming claims. This is among the vital situation in any organization where the focus is to ensure the authentication of user while detecting the fraud behaviour of user among the others.

In past several threats have pruned which has led to special problems, especially the Eavesdropping report of patient health records this tends to unauthorized access of communication in layer of network where the attacker tries to illegally sniffs into the communication network and detect the patient data with unlawful interception. Another security breach known as man-in the middle threat is commonly known attack the usually occurs where the two communicating networks were breach with third party and access is gain between information channel and attacker gain the access of entire data flow in the communication protocol. To deal with threats, endpoint authentication processes are determined which include cryptographic protocols.

2. Encrypting Data

The data encryption means the overall orientation of data is encrypted to minimize the breach of security in data flow. In healthcare organization the data usage is from healthcare practioners to patient and hospital, so the devices are connected to each other via network. The encryption of data can efficiently reduce the packets sniffed and minimize threat. Further, the keys hold by each node should be minimized to reduce the causes privacy and security breach. In past several, algorithms are developed for encryption but the success in big data is still a feasible study of approach.

3. Integrity

The integrity of data should be maintained as the information transferred should not be modified by the attacker. In general, the attacker modifies the original value with some modified values. This is the most popular attack on the data where the frisking is done with personal data of user which may include social security number, data of birth, address of the users and other values. Several data anonymization techniques are discussed which include k-anonymity to protect the values being replaced with modified values. But these methods suffer varied drawbacks in big databases so, a significant technique needs to developed and deployed by overlooking the need of privacy and security needs.

4. Auditing

The secure data auditing is required to depict the security and privacy breach in the network or any intrusion detection. Auditing can generalize the user activity while identifying the log records in healthcare databases, so as to detect any modification or access of data. Several studies in past for intrusion detection are recorded to measure the traffic flow or data flow. Hence, solution exists if the security breach exists then the data was stored in distributed network for ensure the healthcare system. In, context to big data flow via network the system should be able to find abnormalities flowing in the network and should substantiate the alerts in heterogeneous environment. Hence, several integrated frameworks are discussed for deployment in real world scenario.

5. Availability

The data must be provided to legitimate user whenever required and any delay information can affect overall patient diagnosis and can lead to clinical implications. But the control and access of should be vitrified to the authentic user and the control policy should also be governed with prioritized to users' access. This, system can ensure the privacy of patient requirements where the specific privilege permissions are granted to users where the control is at administrator end.

3 A Novel Framework for Privacy in Big Data Analytics

The healthcare organization are facing challenge in day to day scenario for manging their data and safeguarding it from cyber-attacks. This growing need of data privacy and knowledge discovery from big databases is a challenging task where the focus is to generate

patterns for future diagnosis and prognosis of disease with privacy and security. In the current approach we have applied the big data analytic techniques for patients suffering from Tuberculosis (TB) and HIV while keeping in concern with maintaining privacy of patient personal data. The current approach of study is focused on to detect patterns from healthcare databases for future decision making. The novel framework is designed to capture data from varied heterogeneous resources for patterns discovery. Figure 1 represents the framework for retrieval of information and decision making.

3.1 Data Capture

In this study, the entire population of US in all the 50 states was taken into consideration which was around 323.1 million. The population mainly involved people working in service sector.US population involves people of different races like Hispanic, Non-Hispanic, African American, Asian, Native Hawaiian, American Indian, Latino, White, Pacific Islanders. US economy is considered as a Developed Economy by all the categorising organisation be it World Bank or United Nations (UN). Data was obtained from OTIS (Online Tuberculosis Information System), a data repository of CDC (Centre for Disease Control), which is a major operating component of Department of Health and Human Services. It conducts extensive research and provides valuable information about different health issues. The data obtained consists of the information of 1,79,625 patients which was further subdivided into different categories which includes: Year wise: The data contained information of TB patients from year 1993–2014; Age Group wise: The database was further classified based on their age into different groups: 0–4, 5–14, 15–24, 25–44, 45–64, 65+ years of age; Race/Ethnicity wise: The data of TB patients to be studied was classified according to race/ethnicity of patients under subcategories: “Asian, Non-Hispanic”, “Multiple Race, Non-Hispanic”, “American Indian or Alaska Native, Non-Hispanic”, “White, Non-Hispanic”, “Native Hawaiian or Other Pacific Islander, Non-Hispanic”, “Black or African American, Non-Hispanic”, “Hispanic or Latino”; HIV status wise-Data also contained information about the HIV status of the patients; Socio-economic Practices wise:

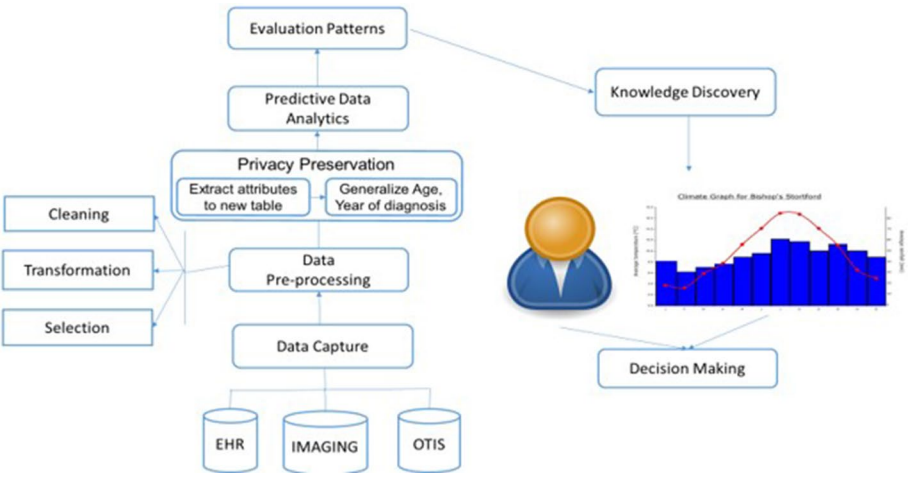

Fig. 1 Novel framework for privacy in big data analytics

The data contained information about whether the TB patients consumed alcohol or other types of drugs(injecting or non-injecting); Vital Status: The data also contained the status of the TB patients whether the patients are alive or died due to TB.

3.2 Data Pre-processing

Data pre-processing is among the major step for evaluation of patterns. As real time datasets consist of missing values, noisy values and inconsistent data records which needs to handle effectively for diagnosis and prognosis of disease. Data pre-processing comprises of varied steps which include data cleaning, data transformation, data selection, data integration and others. Hence, the steeps are pathways for accessing high quality which can deliver imperative results for future prediction of disease. However, after applying data pre-processing dataset is prepared to be utilized for further investigation of study. In current approach, we have applied data cleaning, data transformation and data selection for discovery of knowledge from large scaled database [25].

Data Cleaning The real-world datasets comprise of missing values and noisy values which can generate bungling results thereafter effecting overall decision making. Hence, data cleaning is applied for removing missing values and replacing them with mean values for retrieval of effective and efficient patterns for knowledge discovery process. Thus, applied step was examined for varied techniques which include missing values replaced by NULL values, manually entering the values which is again a very time-consuming step finally the most appropriate technique was removing the value with mean or median which was considerable benefited technique to detect relevant results for knowledge discovery.

Data Transformation It is the technique to transform the data with varied scale ranges from 0.0 to 1.0 in respective of data values. The transformation is accomplished for several techniques from classification, neural networks and clustering for normalizing data sets values. There are several well-known data transformation techniques which Z-Score for normalization of data are, decimal scaling for scaling with varied decimal ranges and min–max normalization. In the current approach, we have utilized min–max technique for transformation of data.

Data Selection The feature selection is binding step for identifying the most appropriate features which correlate among each other and removing irrelevant and superfluous attributes. Thereby, the dimensionality of data is reduced which can severely increase the optimization technique of algorithm for discovery of patterns and knowledge. Several techniques such gain ration, information gain, Correlation based techniques are applied to determine the most effective attributes for knowledge discovery. However, we have applied correlation-based technique for data transformation as the features were more correlated with class values as compared to each other and no other technique was imposing better results.

3.3 Medical Data and Privacy Preserving Data mining

When we talk about medical data specially the patient health data, there is a big requirement of privacy to safeguard a patient's details. The medical records can be snooped by insurance companies, medical laboratories or advertising firms to franchise their commercial interests. Since, clustering data mining technique does not rely on class labels, it forms the first choice of data analytics method in conjunction with privacy preservation. The patient attributes need not be mapped to specific class labels for predictive analysis. When we talk about a patient, the person can be identified by primary attributes such as,

name, age, social security number, or a set of secondary attributes like occupation, history of diagnosis and so on. These attributes together make a dataset which can uniquely identify a person. To provide privacy to the patient, we have tested a hybrid solution which comprises of vertical partitioning of patient attributes and anonymization through generalization of some specific attributes, for instance generalization of age into a range of age groups, and finally analysis through K-means clustering.

3.4 Predictive Data Analytics

The data analytics techniques are the major step for decision making process among the large-scale databases. It tends to identify the process where the algorithmic powers are utilized for predictive modelling. However, determine the appropriate data mining technique for discovery of patterns among large scale databases is substantial for decision making process. Data mining techniques are indispensably distributive in two categories descriptive and predictive techniques. In the current approach we have utilized the predictive technique to discover clusters of variable size and number. The overall results were retrieved by Matlab 7.1 and STATA based software. The K-Means clustering technique was applied to retrieve deterministic results for future decision making. The K means clustering works on the principle to redefining mean values for each cluster using K as number of clusters. The centres must be chosen very thoughtfully as different location of centres leads to different results. The best way is to place these centres as far away from each other as possible. Next step involves considering each point or value of the dataset and linking it to the nearest centre. After associating all the points to the centres primary grouping is said to be completed. Further, new centroids are defined from the clusters obtained in previous step. After new k centroids are defined the same data points are linked to new centroids. This step is repeated again and again till no changes are done or centres do not move further.

$$M(O) = \sum_{i=1}^{D} \sum_{j=1}^{Di} \left\| (p_i - n_j) \right\|^2$$

$\left\| (p_i - n_j) \right\|^2$ = Euclidean distance between p_i and n_j.
D_i = number of data points in ith cluster.
D = number of cluster centres.

Simple k means clustering algorithm clusters data depending upon different distance-based methods which include Euclidean Distance based method, Manhattan Distance based Method, Chebyshev Distance based Method, Filtered Distance based Method and Minkowski Distance based Method. Each of the distance methods have different ways to calculate distances between two points. Thus, each distance method gives different clusters where the quality of clusters obtained differ for each variable dataset.

Knowledge Discovery and Decision making: It tends to be the final process where the discovery of knowledge can be interpreted with the prediction of results for future intervention policy or decision making. The outcome of research proves that the identified model is capable to determine the realistic knowledge or unable to achieve the probabilistic results. Hence, identifying the retrospective factors which can be enabled for iterative and judgmental decision making. If the results tend to be incompatible or inconsistent then the recursive process can be synthesized for decision making.

Table 1 Precision accuracy of extracted attributes

Attribute	Precision value
Year of diagnosis	1
Age	0.992
Race	1
HIV status	0.995
Vital status	0.97

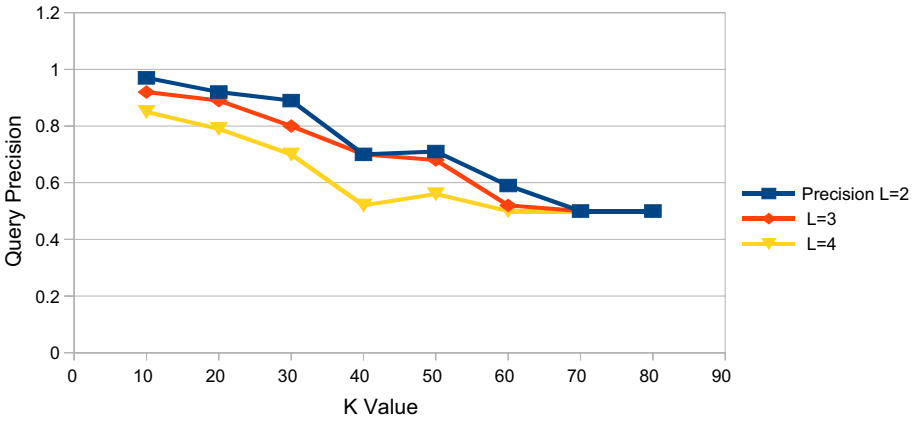

Fig. 2 Precision at different levels of privacy anonymization

4 Results

The data was collected for HIV and TB patients to determine the overall correlated patterns which are the root cause for prognosis of disease. The data obtained consists of evidences among 1,79,625 patients which were further correlated with socio economic factors which included race, age, gender, HIV status, and others from year 1993 to 2014 [23, 24, 40].

To implement privacy preservation, the dataset to be analysed was first partitioned vertically. The attributes which were found relevant for clustering were year of diagnosis, age, race, HIV status and vital status. The attribute year of diagnosis was generalised under the ranges 2000–2002, 2003–2005 and 2006–2010. The attribute age was generalised under the age groups 15–24, 25–44, 45–64 and 65+, which were also in conjunction with the clustering strategy.

To evaluate the weight of the shortlisted attributes, we computed the precision value given by, $\text{Precision} = C_i/T$, where C_i is the measure of correctly extracted information and T is the total information that was extracted without any partitioning. The precision values are shown in Table 1.

To test and compare the K anonymization, the attributes were queried with different levels of generalizations. The Fig. 2 shows varying levels of precision in the query results for anonymity threshold K, where $0 \leq K \leq 90$. As we can see, the precision of results falls sharply for high levels of anonymization, and thus the information loss. For maximal results, we have limited anonymity level to an optimal level of $L = 2$.

The data was then analysed for HIV and Tuberculosis co-infection to determine the interrelated factors for discovery of knowledge. The prognosis of disease was measured with number of occurrences with total population in consideration. Similar, approach was synthesized for the prognosis of TB with HIV to calculate patient suffering from both HIV and TB, proportional technique was involved by dividing the prognosis of both TB and HIV incidence with total population in consideration. Relative incidence rate of TB and HIV with respect to TB without HIV was calculated by diving the incident rates of the above two. Mortality rates were calculated for all the above variables in the same way. Finally using this refined data, Classification models were constructed using J-48 decision tree. The data was clustered into different clusters.

In Table 2 variable clusters of varied shapes and sizes were determined, where 4 clusters were obtained through Simple k means clustering algorithm utilising Euclidian Distance Based method. Cluster0 contained mostly contained population between 45 and 64 years of age, Hispanic or Latino race and HIV status was considered negative with most of them being alive. Cluster 1 contained population of Age group 65+, white race and most of them were dead. Cluster 2 consists of 25–44 years of age with HIV status negative and most of them were also alive. Finally, cluster 3 contained of 15–24 years of age which had HIV status as negative and were also alive.

About 1,79,625 Tuberculosis cases were reported in United States from year 1993 to 2014, out of which 18% (i.e. 32,636) were diagnosed with HIV. Incidence of TB cases initially increased from 1993 (7329 cases) to 1996 (8744 cases) which is about 19.3% increase. From Year 1996 to 2000 TB cases showed a decline of about 7.6% with 8072 cases reported in year 2000. From Year 2000 to 2004 TB cases incidence increased by about 5% with 8476 cases reported in Year 2004. In Year 2004–2009 the trend was in lower side where TB cases decreased by 12.8% with only 7383 cases of TB reported in Year 2009. In 2009–2011 witnessed a huge surge in TB cases with an astonishing 18.9% increase with 8784 cases being reported in Year 2011. From year 2012 to 2014 TB cases declined gradually with only 8173 cases reported in year 2014 which is a 6.9% decrease. Figure 3 represents the overall clusters retrieved.

In Fig. 4 overall increase in the prognosis of TB from 1993 to 2014 is measured using STATA where the data represents the step high in year 2011. Further, the study was implemented to measure the age wise trends in TB patients' year wise. The results represented that different age groups have different susceptibility towards TB disease with a clear difference in the number of cases reported for each of different age groups. Age group 0–4 years consisted of only 2% TB cases reported in US from year 1993 to 2014. Age

Table 2 Clusters obtained Simple k Means clustering analysis

Clusters	Cluster0	Cluster1	Cluster2	Cluster3
Year	2002	2003	2006	2000
Age	45–64	65+	25–44	15–24
Race	Hispanic or Latino	White, Non-Hispanic	American Indian or Alaska Native, Non-Hispanic	African American, Non-Hispanic
HIV status	Negative	Negative	Negative	Negative
Vital status	Alive	Dead	Alive	Alive
Value	146.10	15.79	129.42	83.71

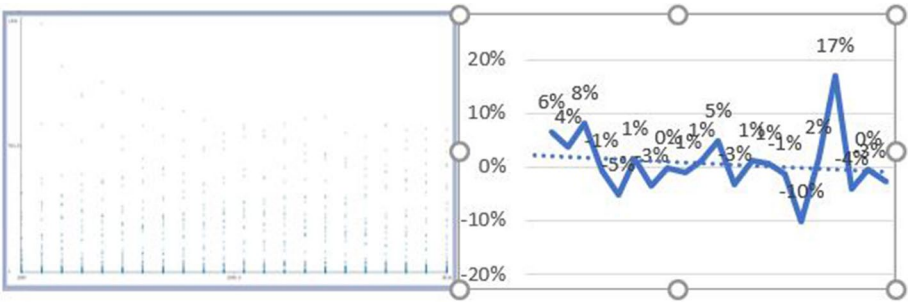

Fig. 3 Number of TB cases in each year

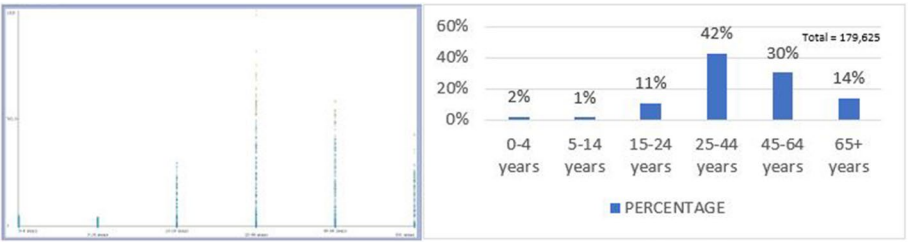

Fig. 4 Clusters correlated age and year wise

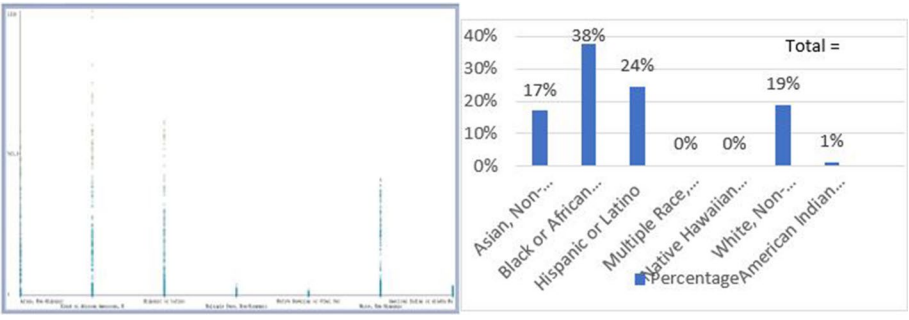

Fig. 5 Clusters of TB cases in each race

Group 5–14 years represented even lesser proportion as only 1% TB patients out of all the TB patients belonged to this age group. Age group 15–24 years that is young adults comprised of about 11% of all the TB patients. Age group 25–44 years i.e. Middle-aged population comprised of the maximum proportion of TB patients with almost 42% of all the reported TB cases belonging to this age group. Age Group 45–64 Years of age i.e. old people comprised about 30% of all the TB cases. People with 65+ age comprised 14% of the TB cases.

In Fig. 5, the graph represents the high percentage of patients suffering in age group of 25–44 where the substantial policies should be determined at managerial level low down the prognosis of disease.

Further, the data was analysed for races that are inhabiting US discriminately with some races getting affected with TB more as compared to other races. The results obtained represented that Asian, Non-Hispanic, Black or African American, Hispanic or Latino were observed to suffer from TB supplementary rater than as compared to Multiple Race, Non-Hispanic, Native Hawaiian or other Pacific Islander, Non-Hispanic, White, Non-Hispanic. The American Indian or Alaska Native, Non-Hispanic were suffering with lower rate as compared to others. In Fig. 4 an illustration is presented race wise.

In Fig. 6 the socio-economic trends were depicted with drug usage of TB patients. The results obtained clearly represented that drug use (injecting or non-injecting) or alcohol use does not have significant effect on the vital status of the TB patient as the graph obtained from analysis of clusters did not show any clear patterns to prove the above stated. The distribution of those dying and those who were alive was completely random throughout all the years.

Similarly, the focus of study was to determine the correlated factors of TB with HIV and their socio-economic causes of dead and alive. Thus, further the study was evaluated to determine the synchronized patterns for knowledge discovery among large scale databases. The outcome of the study represented that there exists a definite trend with number of HIV–TB cases from Year 1993 to 2014. The data exhibited a regular decline with the number of HIV–TB patients from year 1993 to 2014. Hence, in year 1993, HIV–TB patients comprised about 50% of all the TB infected patients. But with each passing year number of HIV–TB patients kept on decreasing in a regular manner with only 6% HIV–TB patients in Year 2014. Table 3 represents the overall rate of TB and HIV patients with their alive and dead status year wise.

Likewise, Fig. 7 represents the cluster results year wise with total HIV and NON-HIV patients suffering from TB year wise.

However, mortality rates in HIV and TB patients were observed to be high as compared to Non-HIV TB patients. As data represents the mortality rate was 6% higher with patients suffering from both HIV and TB as compared to only 1% death rate in Non-HIV TB patients. In Fig. 8, year wise cluster were rationalized where 2 clusters were determined with dead and alive status year wise, cluster0 epitomizes the patients which were alive and cluster1 with dead percentage year wise. The results predicted that with each passing year death rates in HIV–TB patients declined in a linear fashion with only 2% death rate recorded in year 2014. While that of Non-HIV TB patients declined from 2% in year 1993 to 1% in year 2014.

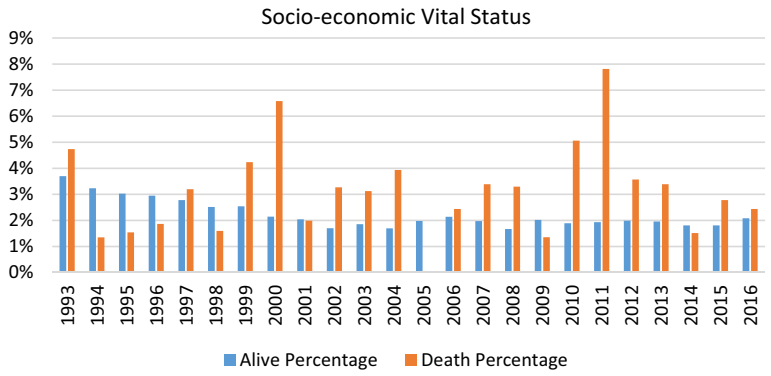

Fig. 6 Vital status of TB patients with drug use and alcoholism in each year

Table 3 Percentage of HIV TB and non HIV TB cases along with vital status of both scenarios

Year	Total negative (%)	Alive (%)	Dead (%)	Total positive (%)	Alive (%)	Dead (%)
1993	50	99	1	50	94	6
1994	54	99	1	46	95	5
1995	63	99	1	37	95	5
1996	70	99	1	30	94	6
1997	76	99	1	24	96	4
1998	78	99	1	22	96	4
1999	79	99	1	21	96	4
2000	82	99	1	18	95	5
2001	83	99	1	17	97	3
2002	83	99	1	17	97	3
2003	84	99	1	16	97	3
2004	86	99	1	14	96	4
2005	87	99	1	13	98	2
2006	88	99	1	12	96	4
2007	89	99	1	11	96	4
2008	90	99	1	10	97	3
2009	90	100	0	10	96	4
2010	92	99	1	8	95	5
2011	92	99	1	8	97	3
2012	93	99	1	7	97	3
2013	93	99	1	7	97	3
2014	94	100	0	6	98	2

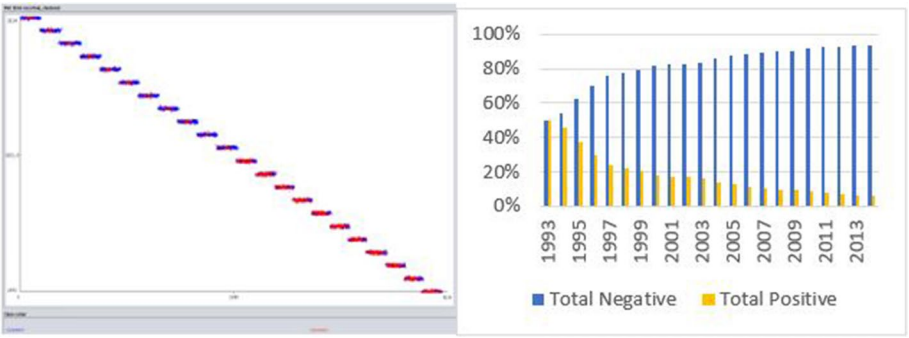

Fig. 7 HIV and non-HIV TB cases

In Fig. 9, statistical representation of HIV with TB patients determines the decrement in mortality rate from 1993 to 2014 where it has drastically reduced from 6 to 2%.

In Table 4, a trend was depicted for patients suffering from both HIV and TB in correspondence to variable age groups. The results suggested a high percentage of HIV TB cases belonged to two categorize of age groups i.e. 25–44 (comprising of middle aged

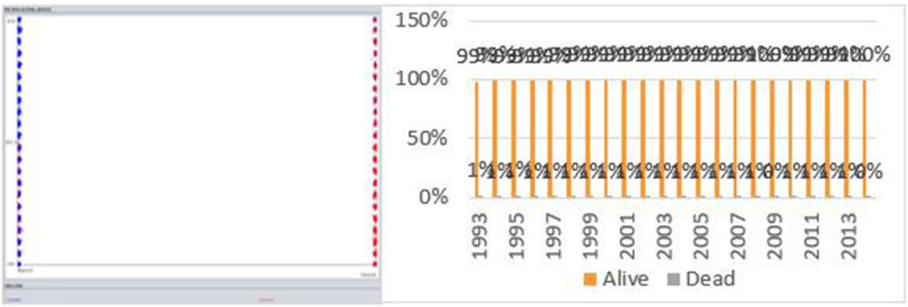

Fig. 8 Cluster year wise vital status in HIV and non HIV TB cases

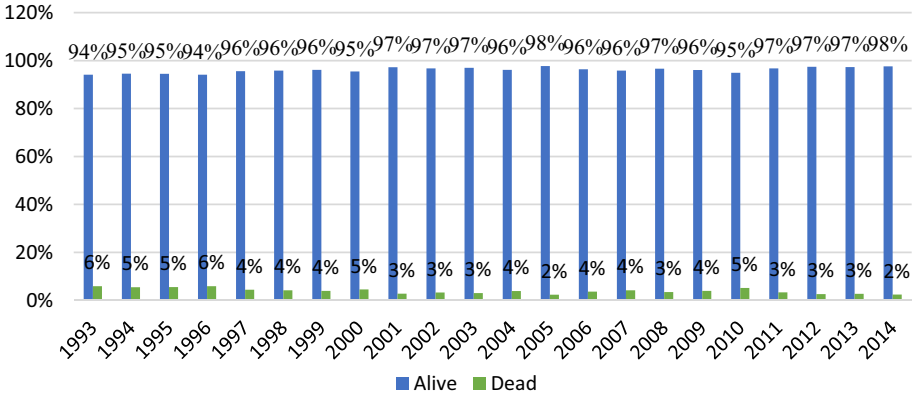

Fig. 9 Year wise vital status of HIV TB cases

Table 4 Age wise % of HIV and non-HIV TB cases with vital status

Age	Total negative (%)	Alive (%)	Dead (%)	Total positive (%)	Alive (%)	Dead (%)
0–4 years	96.8	99.8	0.2	3.2	96.8	3.2
5–14 years	96.8	100.0	0.0	3.2	98.8	1.2
15–24 years	94.9	99.8	0.2	5.1	98.3	1.7
25–44 years	71.8	99.6	0.4	28.2	95.8	4.2
45–64 years	82.8	99.1	0.9	17.2	95.1	4.9
65 + years	97.2	97.9	2.1	2.8	90.7	9.3

adults) having 28.2% of HIV TB cases out of total HIV TB cases and 45–64 age group (comprising of old people) having 17.2% of HIV TB cases out of total HIV TB cases.

In Fig. 10, the cluster analysis represents that age group effected in maximum capacity is 25–44 years with HIV–TB cases.

Additional, results were represented for age wise, vital status for both patients suffering from HIV–TB with Non-HIV TB cases. However, data suggested to have lower mortality rate with patients suffering due to TB. But mortality rates were higher due to TB in

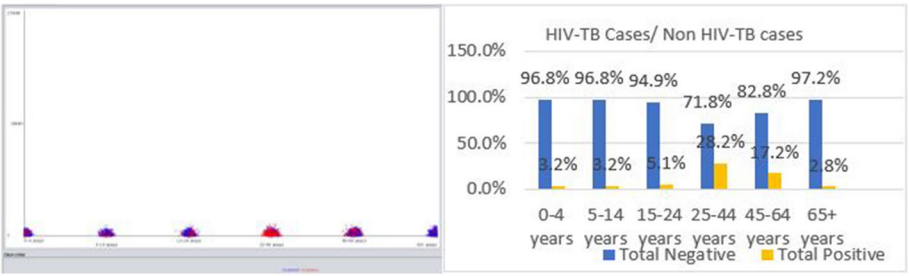

Fig. 10 Cluster age wise total HIV and non HIV TB cases

different Age groups varied with screening higher of mortality rate in 65+ age group for both HIV TB (mortality = 9.3%) and Non-HIV TB (mortality = 2.1%) cases whereas lower mortality rates were observed in age between 5 and 14. In general HIV TB patients showed a much higher mortality as compared to Non-HIV TB patients in all the age groups. Figure 11, represents two clusters where cluster0 blue in colour represents age wise status for dead, cluster1 represents dead status age group wise.

The statistical measure of Non-HIV TB cases age wise with vital status of dead and alive. There is considerable rise in mortality rate in the age of 65+.

In Fig. 12, data perceives to be higher in mortality among the all age groups discussed. However, age after 45 years is more probable to have higher mortality rate as compared to other age groups.

In Table 5, a Discriminant analysis is represented where the incidence of HIV–TB is significant with different races living in US with some races getting affected with HIV TB more often as compared to other races. As observed from the results obtained from data analysis HIV TB is more common in "Black or African American" (29.3%), "Hispanic or Latino" (17.1%), "White Non-Hispanic" (12.8%). Whereas it is much less common in "Asian, Non-Hispanic" (2.6%); "Multiple Race, Non-Hispanic" (5.0%); "Native Hawaiian or Other Pacific Islander, Non-Hispanic" (2.6%); "American Indian or Alaska Native, Non-Hispanic" (5.3%).

In Fig. 13, clustering analysis were performed for different races to measure the mortality rate among HIV–TB and Non-HIV TB cases. Cluster0 blue in colour represents

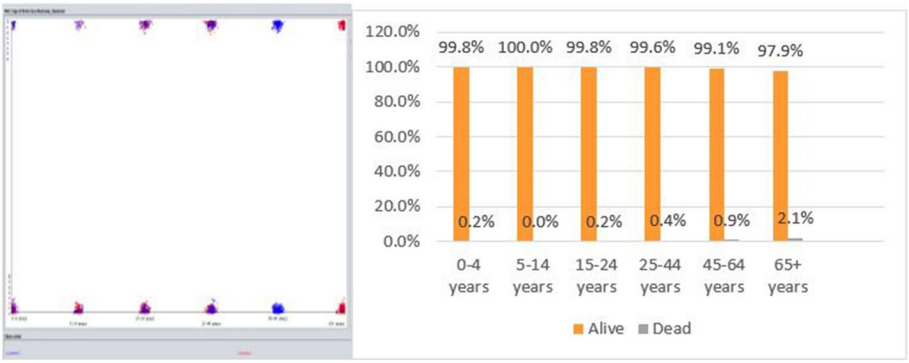

Fig. 11 Cluster age, vital status in HIV and Non HIV TB cases

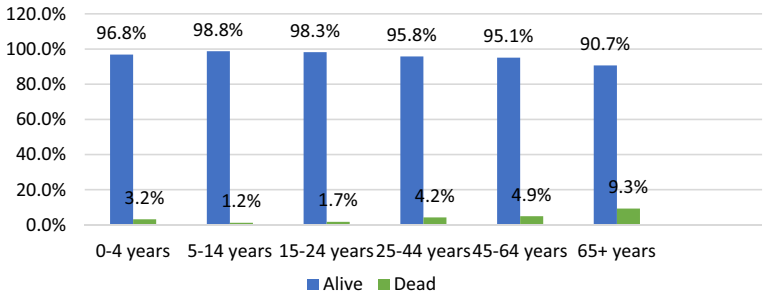

Fig. 12 Age wise vital status of HIV TB cases

age wise status for alive, cluster1 represents dead status race wise. Results obtained after analysis of TB cases in different races and their survival rates revealed that different races showed different mortality due to the disease. As, previous results were suggestive of higher mortality rates for patients suffering HIV–TB cases, similarly also the HIV TB patients showed higher mortality as compared to Non-HIV TB. The race wise analysis also showed that "American Indian or Alaska Native, Non-Hispanic" Race showed maximum mortality (1.5%) whereas "Asian, Non-Hispanic Race" and "Native Hawaiian or Other Pacific Islander, Non-Hispanic" Race showed minimum mortality (0.3%) in case of Non-HIV TB cases.

In Fig. 14, HIV TB cases indicated different trends in mortality rates all together where "White, Non-Hispanic" showed maximum mortality (5.3%) followed by Black or African American, Non-Hispanic Race (4.6%), "American Indian or Alaska Native, Non-Hispanic" Race (4.4%) and "Hispanic or Latino" Race (3.7%). Whereas "Native Hawaiian or Other Pacific Islander, Non-Hispanic" Race showed minimum mortality (0%) followed by "Asian, Non-Hispanic" Race (2.4%) and "Multiple Race, Non-Hispanic" (2.8%).

5 Conclusion

Big data has its potential opportunities in healthcare application areas due to faced IT based technological interventions. The technical challenge exists due to impeding obstacles which occurs owing lapses in security and privacy of data. In this paper, we discuss the big data analytics with privacy and security in context with healthcare databases for patients suffering from HIV and TB. The paper briefly discusses the related works across the big healthcare databases in context with predictive data analytics with privacy and security. The current approach of study is focused on to detect patterns from healthcare databases with concerns on maintain the privacy of data and generating the patterns for future clinical decision making. Further, the study was intuitive and detect patterns which can generate knowledge while maintaining the privacy among the data so no person information was instinctively analysed to generate patterns.

A novel framework was designed to effectively and efficiently capture big data from various resources in context to maintain the privacy and security of data and detect hidden patterns for clinical decision making. In general, focus of study is impacted on big data analytics as an imperative technology to generate impounded outcomes to reduce the global burden of disease and preserve the patient personal information. Hence, we found that big

Table 5 Race, vital status of total HIV and non-HIV TB cases

Races	Total negative (%)	Alive (%)	Dead (%)	Total positive (%)	Alive (%)	Dead (%)
"Asian, Non-Hispanic"	97.4	99.7	0.3	2.6	97.6	2.4
"Black or African American, Non-Hispanic"	70.7	99.0	1.0	29.3	95.4	4.6
"Hispanic or Latino"	82.9	99.5	0.5	17.1	96.3	3.7
"Multiple Race, Non-Hispanic"	95.0	99.0	1.0	5.0	97.2	2.8
"Native Hawaiian or Other Pacific Islander, Non-Hispanic"	97.4	99.7	0.3	2.6	100.0	0.0
"White, Non-Hispanic"	87.2	98.9	1.1	12.8	94.7	5.3
"American Indian or Alaska Native, Non-Hispanic"	94.7	98.5	1.5	5.3	95.6	4.4

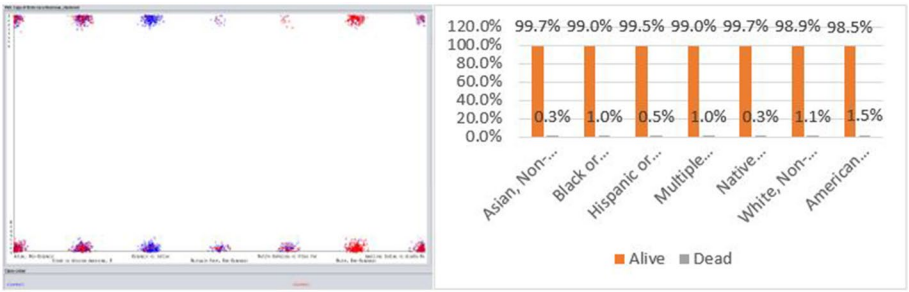

Fig. 13 Cluster with race, vital status of HIV and non-HIV TB cases

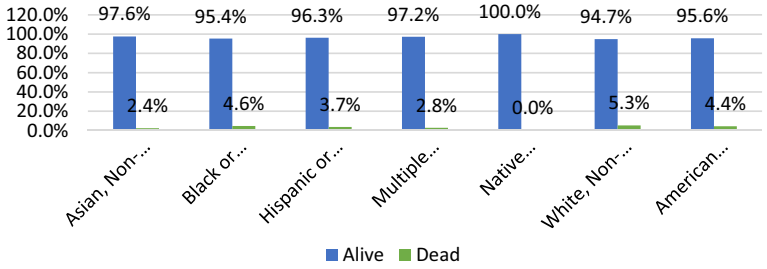

Fig. 14 Vital status of HIV and TB cases

data analytics techniques bound to have extensive vision where the paradigm to generate and develop sustainable for varied application domains in which the focus of study is influenced for healthcare databases. In technological context there exist huge advancement in electronic databases with volume and complexity, herewith the knowledge discovery is discovered as an exponential tool to analyse big data.

The objective of this study was to observe the big data analytical trends of patients suffering from TB–HIV in US population from 1993 to 2014 and further enabling the techniques to preserve the data privacy at each level, so empowering a secure platform for knowledge discovery. Thus, we observed the patients correlated patterns keeping in concern without hampering the personal data of patients hence, the personal information was hidden and other associated patterns were analysed with socio-economic, age groups and races inhabiting and thus seek to understand HIV–TB syntomic cause for prognosis of disease. The results attained clearly exhibited that incidence levels in different years has no uniform trend. Though a clear trend was observed in incidence levels in different age groups as age group 25–44 and 45–64 were the most infected of all the ages. Race wise analysis represented that "Black or African American" and "Hispanic or Latino" showed maximum incidence as compared to other races in US.

HIV–TB coinfection analysis suggestive that the HIV–TB coinfection has decreased significantly from year 1993, where it was 50% of total TB cases to year 2014 where HIV–TB coinfection cases were only 6% of all TB cases. Though most of the patients infected from TB received treatment in time and survived, but results also evidently exhibited that when HIV–TB coinfection occurred it caused more mortality as compared to non-HIV TB cases. It was also inferenced from the results that the mortality in HIV TB cases also declined from 1993 to 2014.

Acknowledgements This research work is catalyzed and supported by Indo-Polish joint research grant DST/INT/POL/P-02/2014 and National Council for Science and Technology Communication (NCSTC) research grant 5753/IFD/2015-16 funded by Department of Science and Technology (DST), Ministry of Science and Technology (Govt. of India), New Delhi, India [Grant recipient: Dr. Harleen Kaur].

References

1. Xu, L., Jiang, C., Wang, J., Yuan, J., & Ren, Y. (2014). Information security in big data: Privacy and data mining. *Journal of Rapid Open Access Publication, 2,* 1149–1176.
2. Yu, W. D., Kollipara, M., Penmetsa, R., & Elliadka, S. (2013). A distributed storage solution for cloud based e-Healthcare Information System. In *Proceedings of the IEEE 15th international conference on e-health networking, applications & services (Healthcom'13); Lisbon, Portugal* (pp. 476–480).
3. Athey, B. D., Braxenthaler, M., Haas, M., & Guo, Y. (2013). Transmart: An open source and community-driven informatics and data sharing platform for clinical and translational research. *AMIA Summits on Translational Science Proceedings, 2013,* 6–8.
4. Jeanquartier, F., & Holzinger, A. (2013). On visual analytics and evaluation in cell physiology: A case study. In A. Cuzzocrea, C. Kittl, D. E. Simos, E. Weippl, & L. Xu (Eds.), *Availability, reliability, and security in information systems and HCI* (pp. 495–502). Berlin: Springer.
5. Jiang, M., Zhang, S., Li, H., & Metaxas, D. N. (2015). Computer-aided diagnosis of mammographic masses using scalable image retrieval. *IEEE Transactions on Biomedical Engineering, 62*(2), 783–792.
6. Johnston, M. E., Langton, K. B., Brian Haynes, R., & Mathieu, A. (1994). Effects of computer-based clinical decision support systems on clinician performance and patient outcome: A critical appraisal of research. *Annals of Internal Medicine, 120*(2), 135–142.
7. Jung, K., LePendu, P., Iyer, S., Bauer-Mehren, A., Percha, B., & Shah, N. H. (2014). Functional evaluation of out-of-the-box text-mining tools for data-mining tasks. *Journal of the American Medical Informatics Association, 22*(1), 121–131.
8. Vararuk, A., Petrounias, I., & Kodogiannis, V. (2007). Data mining techniques for HIV/AIDS data management in Thailand. *Journal of Enterprise Information Management*. https://doi.org/10.1108/17410390810842255.
9. Asha, T., Natarajan, S., & Murthy, K. N. B. (2011). A data mining approach to the diagnosis of tuberculosis by cascading clustering and classification. *Journal of Computing 3* arXiv:1108.1045 [cs.AI].
10. Uçar, T., & Karahoca, A. (2011). Predicting existence of Mycobacterium tuberculosis on patients using data mining approaches. *Procedia Computer Science, 3,* 1404–1411.
11. Garg, S., & Rupal, N. (2015). A review on tuberculosis using data mining approaches. *International Journal of Engineering Development and Research, 3*(3), 1–4.
12. Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. *Journal of Parallel and Distributed Computing, 74*(7), 2561–2573.
13. Kawamoto, K., Houlihan, C. A., Andrew Balas, E., & Lobach, D. F. (2005). Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success. *BMJ, 330*(7494), 765.
14. Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics, 8*(1), 1–8.
15. Metcalfe, J. Z., Porco, T. C., Westenhouse, J., Damesyn, M., Facer, M., Hill, J., et al. (2013). Tuberculosis and HIV co-infection, California, USA, 1993–2008. *Emerging Infectious Diseases, 19*(3), 400.
16. Kim, S.-H., Kim, N.-U., & Chung, T.-M. (2013). Attribute relationship evaluation methodology for big data security. In *2013 international conference on IT convergence and security (ICITCS), IEEE* (pp. 1–4).
17. Rama Lakshmi, K., & Prem Kumar, S. (2013). Utilisation of data mining techniques for prediction and diagnosis of major life threatening diseases survivability-review. *International Journal for Scientific and Engineering Research, 4*(6), 923–932.
18. https://www.cdc.gov/about/organization/cio.html.
19. https://wonder.cdc.gov/tb.html.
20. Sánchez, M. A., Uremovich, S., & Acrogliano, P. (2009). Mining Tuberculosis Data. In P. Berka, J. Rauch, & D. A. Zighed (Eds.), *Data mining and medical knowledge management: Cases and applications*. New York: Medical Information Science Reference.
21. Han, W., Susilo, Y., & Yan, J. (2012). Privacy preserving decentralized key-policy attribute-based encryption. *IEEE Transactions on Parallel and Distributed Systems, 23,* 2150–2162.

22. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105). Curran Associates.
23. Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. *Proceedings of the VLDB Endowment, 5*(12), 2032–2033.
24. Lalys, F., Riffaud, L., Bouget, D., & Jannin, P. (2012). A framework for the recognition of high-level surgical tasks from video images for cataract surgeries. *IEEE Transactions on Biomedical Engineering, 59*(4), 966–976.
25. Langs, G., Hanbury, A., Menze, B., & Muller, H. (2013). VISCERAL: Towards large data in medical imaging challenges and directions. In *Medical content-based retrieval for clinical decision support* (Vol. 7723, pp. 92–98). Springer.
26. Yazan, A., Yong, W., & Raj Kumar, N. (2015). Big data life cycle: Threats and security model. In: *21st Americas conference on information systems.*
27. Greenleaf, Graham and Chung, Philip and Mowbray, Andrew, Influencing Data Privacy Practices By Global Free Access: The International Privacy Law Library (November 14, 2014). UNSW Law Research Paper No. 2014-56.
28. OECD. (2013). *Data-driven healthcare innovation, management and policy, DELSA/HEA(2013) 13*. Paris: OECD.
29. Chauhan, R., & Kaur, H. (2017). A feature based reduction technique on large scale databases. *International Journal of Data Analysis Techniques and Strategies., 9*(3), 207–221.
30. Chauhan, R., Kaur, H., & Chang, V. (2017). Advancement and applicability of classifiers for variant exponential model to optimize the accuracy for deep learning. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-017-0561-x.
31. Kaur, H., Chauhan, R., & Wasan, S. K. (2014). A Bayesian network model for probability estimation. In M. Khosrow-Pour (Ed.), *Encyclopaedia of information science and technology* (3rd ed.) (pp. 1551–1558). Retrieved December 10, 2014, from https://doi.org/10.4018/978-1-4666-5888-2.ch148.
32. Chauhan, R., & Kaur, H. (2015). Big data application in medical domain. In D. P. Acharjya, et al. (Eds.), *Computational intelligence for big data analysis: Frontier advances and applications. Volume 19 of the series adaptation, learning, and optimization* (pp. 165–179). Basel: Springer.
33. Kaur, H., Tao, X. (2014). ICT and Millennium Development Goals: A United Nations Perspective, pp. 271, Springer, New York.
34. Chauhan, R., Kaur, H., Lechman, E., Marszk, A. (2017). Big data analytics for ICT monitoring and development. In: Kaur, H., et al. (eds.) Catalyzing Development Through ICT Adoption: The Developing World Experience, pp. 25–36. Springer, New York.
35. Hu, P., & Gao, H. (2017). A key-policy attribute-based encryption scheme for general circuit from bilinear maps. *International Journal Network Security, 19*(5), 704–710.
36. Lai, J., Deng, R. H., Guan, C., & Weng, J. (2013). Attribute-based encryption with verifiable outsourced decryption. *IEEE Transactions on Information Forensics and Security, 8*(8), 1343–1354.
37. Lee, C. C., Chung, P. S., & Hwang, M. S. (2013). A survey on attribute-based encryption schemes of access control in cloud environments. *International Journal Network Security, 15,* 231–240.
38. Lewis, G., Echeverria, S., Simanta, S., Bradshaw, B., & Root, J. (2014). Tactical cloudlets: Moving cloud computing to the edge. In *IEEE military communications conference* (pp. 1440–1446).
39. Li, J., Huang, X., Li, J., Chen, X., & Xiang, Y. (2014). Securely outsourcing attribute-based encryption with checkability. *IEEE Transactions on Parallel and Distributed Systems, 25*(8), 2201–2210.
40. Agarwal, S., Nguyen, D. T., Teeter, L. D., & Graviss, E. A. (2017). Spatial-temporal distribution of genotyped tuberculosis cases in a county with active transmission. *BMC Infectious Diseases, 17,* 378.
41. Kriegel, H.-P., Kroger, P., & Zimek, A. (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data, 3*(1), 1–58. https://doi.org/10.1145/1497577.1497578.
42. Li, J., Yao, W., Zhang, Y., Qian, H., & Han, J. (2017). Flexible and fine-grained attribute-based data storage in cloud computing. *IEEE Transactions on Services Computing, 10*(5), 785–796.

Dr. Ritu Chauhan is an Assistant Professor at the Amity University, Noida, India. Her key research areas include information analytics, applied machine learning and predictive modeling. *She* is the *author* of various publications and has author *of several reputed books*. She is a member to international bodies and is a member of editorial board of international journals on data analytics and machine learning.

Dr. Harleen Kaur is a faculty member and Distinguished Scientist at the School of Engineering Sciences and Technology at Jamia Hamdard, New Delhi, India. She recently worked as Research Fellow at United Nations University (UNU) in IIGH-International Centre for Excellence, Malaysia to conduct research on funded projects from South-East Asian Nations (SEAN). She is currently working on an Indo-Poland bilateral international project funded by the Ministry of Science and Technology, India, and the Ministry of Polish, Poland. In addition, she is working on a national project, catalyzed and supported by the National Council for Science and Technology Communication (NCSTC), the Ministry of Science and Technology, India. Her key research areas include data analytics, big data, applied machine learning and predictive modeling. She is the author of various publications and has authored/edited several reputed books. She is a member of various international bodies and is a member of the editorial board of international journals on data analytics and machine learning. She is the recipient of Ambassador for Peace Award (UN Agency) and honors and is funded researcher by external groups.

Prof. Dr. Victor Chang is currently a Full Professor of Data Science and Information Systems at the School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, UK, since September 2019. He was a Senior Associate Professor, Director of Ph.D. (June 2016–May 2018) and Director of MRes (Sep 2017–Feb 2019) at International Business School Suzhou (IBSS), Xi'an Jiaotong-Liverpool University (XJTLU), Suzhou, China, between June 2016 and August 2019. He was also a very active and contributing key member at Research Institute of Big Data Analytics (RIBDA), XJTLU. He was an Honorary Associate Professor at University of Liverpool. Previously he was a Senior Lecturer at Leeds Beckett University, UK, between Sep 2012 and May 2016. Within 4 years, he completed Ph.D. (CS, Southampton) and PGCert (Higher Education, Fellow, Greenwich) while working for several projects at the same time. Before becoming an academic, he has achieved 97% on average in 27 IT certifications. He won a European Award on Cloud Migration in 2011, IEEE Outstanding Service Award in 2015, best papers in 2012, 2015 and 2018, the 2016 European award. He is a visiting scholar/Ph.D. examiner at several universities, an Editor-in-Chief of IJOCI and OJBD journals, Editor of FGCS, Associate Editor of TII and Information Fusion, founding chair of two international workshops and founding Conference Chair of IoTBDS and COMPLEXIS since Year 2016. He is the founding Conference Chair for FEMIB since Year 2019. He published 3 books as sole authors and the editor of 2 books on Cloud Computing and related technologies. He gave 18 keynotes at international conferences. He is widely regarded as one of the most active and influential young scientist and expert in IoT/Data Science/Cloud/security/AI/IS, as he has experience to develop 10 different services for multiple disciplines.