EXPLORING MEDICAL DATABASE AND DATA MINING IN THE AGE OF BIG DATA

[#1]Dr. Sk.Yakoob, Associate Professor,
[#2]V.Sai Rama Krishna, Assistant Professor,
[#3]G. Rajeswari, Assistant Professor,
Department of Computer Science and Engineering,
SAI SPURTHI INSTITUTE OF TECHNOLOGY, SATHUPALLY, KHAMMAM.

Abstract: Data mining is used to extract meaning from huge databases. Big Data includes the expression, preparation, mining, and analysis of data. Databases are essential to this revolutionary technique of data handling. The branch of computer science concerned with database administration and comprehension is known as database technology. Database analysis necessitates investigation into both the theory and practice of data storage, administration, design, and application. Our seminars on databases and data extraction benefited many medical experts.

KEYWORDS: big data, data mining, database, method, technology

1. INTRODUCTION

Due to the "great information explosion," "Big Data" has acquired importance in business, banking, and healthcare in recent months. The majority of enterprises are engaged in "big data" research and innovation, which is transforming healthcare and medical training. The use of digital data storage and collecting is becoming more common. Big data is becoming increasingly significant as a result of the sheer volume of data created by healthcare businesses. There are numerous examples of firsts in patient therapy in hospitals around the world.

The dissemination of this free work is governed by the Creative Commons Attribution license. Anyone may use, distribute, and copy the work under the conditions of this license.

When it comes to healthcare IT, data can take many forms. In modern medical research, much data is generated and properly employed.

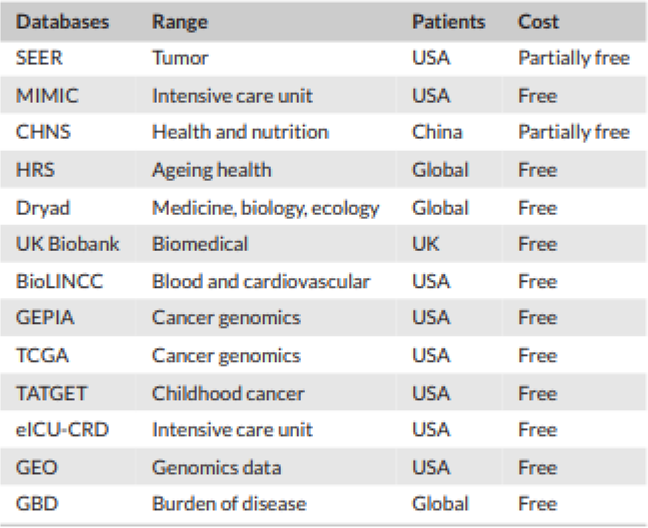
The term "big data" requires further clarification. Because they can process a broader range of data, big data solutions outperform database management systems. Big data has the following characteristics: size, speed, diversity, worth, and veracity8-10. Massive volumes of data must be received and processed rapidly and precisely, which is what we mean by "velocity." "Volume" means "huge in volume," and "velocity" means "speed." The term "variability" is debatable. There are audio, video, and written transcripts available. "Value" certainly lives true to its name, with both a high market value and a small population. When information is correct, meaning is better transmitted. The most difficult component of "big data" is extracting insights from a huge, heterogeneous, and constantly expanding data set. Integrated data analytics is applied in the domains of medicine and molecular biology.

Medical data gathering is difficult due to the broad variety of diseases, treatments, and results. It is difficult to collect information, analyze it, and draw conclusions from it. As we learn more about medicine, we amass more digital data about medical services, health care, and health management. Clinical registration information, administrative claims, electronic health records, biometrics, and patient reports are all examples of "medical big data."15 Big data and data collection are advantageous to healthcare. Diabetes smartphone users are generating critical big data networks through their research, sharing, and communication. Thanks to a US Department of

Health and Human Services effort, big data about patients, providers, and medical records can now be freely shared. Diabetes can be treated with metformin when an EHR examination is completed.

Big data's medical applications are expanding. Medical data can be difficult to organize, which limits access to it. If results need to be altered, the ability to interpret and assess data may be required. The medical literature is vast, current, diverse, incomplete, and critical. Usability, data analysis, and remote coaching are all improved as a result of big data.

Table-1 Check out the NHI Databank.



Databases	Range	Patients	Cost
SEER	Tumor	USA	Partially free
MIMIC	Intensive care unit	USA	Free
CHNS	Health and nutrition	China	Partially free
HRS	Ageing health	Global	Free
Dryad	Medicine, biology, ecology	Global	Free
UK Biobank	Biomedical	UK	Free
BioLINCC	Blood and cardiovascular	USA	Free
GEPIA	Cancer genomics	USA	Free
TCGA	Cancer genomics	USA	Free
TATGET	Childhood cancer	USA	Free
eICU-CRD	Intensive care unit	USA	Free
GEO	Genomics data	USA	Free
GBD	Burden of disease	Global	Free

CHNS stands for Chinese Health and Nutrition Examination. The intensive care unit's medical data library include databases such as the Gene Expression Omnibus, GEPIA, HRS Study, and MIMIC. The terms "Global Burden of Disease" (GBD) and "Surveillance, Epidemiology, and Results" (SEER) are combined.

2. MEDICAL PUBLIC DATABASE OVERVIEW

Every day, our culture generates massive amounts of data. Database science is the study, management, and application of databases. Data analysis includes all areas of database administration. Table 1 contains the most major free medical databases.

Prevalence of disease, diagnostic tests, and ongoing surveillance

SEER was founded by the NCI in 1973 as part of its cancer-fighting efforts. This application can be used to investigate malignant tumors in a variety of ways. SEER's early acceptance was slow.

Medical information mart for intensive care (MIMIC)

Patients in critical care are intensively monitored for organ failure since "extreme medicine" concentrates on catastrophic illnesses and accidents. It is critical to maintain the patient alive and breathing while treating the underlying illness. Some hospital wards are designated for the most seriously ill patients, and both diagnosis and treatment are tracked and assessed. We now have more resources than ever before to research potentially lethal diseases. The application of AI and huge datasets will tremendously assist medical and scientific breakthroughs.

MIMIC was developed by Beth Israel Dikang Medical Center, MIT's Computational Physiology Laboratory, and Philips to aid in critical care research.

The National Institutes of Health funds this essential endeavor. Between 2001 and 2012, clinical diagnostic and treatment data for over 40,000 ICU patients at Beacon Israel Dikang Medical Center were collected. The database's open availability, large sample size, copious data, and continuous patient monitoring all add to its relevance for critical care research. Clinical physicians in the field of severe medicine know a lot, but scientific research lacks standardized clinical diagnosis and treatment data. MIMIC-III 1.4

For the most recent information, go to https://mimic.physionet.org/about/releasenotes/.

The Philips Care Vue Clinical Information System and the IMD Soft Meta Vision ICU System are used to collect patient data. Philips Care Vue was used for patient monitoring from 2001 to 2008, while IMD Soft Meta Vision ICU was used for the next three years (2008-2012).

China health and nutrition survey (CHNS)

The China Resident Health and Nutrition Survey results can be found at http://www.cpc.unc.edu/projects/china. The Chinese Center for Nutrition and Health and the Population Center at the University of North Carolina at Chapel Hill are leading global collaborations to research the 30-year health and nutritional implications of China's family planning policy and socioeconomic transformation. People, families, and neighborhoods are examined economically, demographically, and sociologically. The global survey team includes dietitians, public health experts, economists, sociologists, and demographers. We have you covered from 1989 to the present day, 2012.

In 2015. On June 12, 2018, new data was added to the CHNS database. The data is now shown vertically.

Between 1989 and 2015, ten surveys were investigated. According to the China Health and Nutrition Survey, household income, education level, urbanization, and food policy all have an impact on nutrition, food category consumption, and dietary patterns. Using multistage stratified cluster random sampling, fifteen provinces, autonomous regions, and municipalities in eastern, central, and western China were polled. In August of 2018, around 30,000 people, 7,200 houses, and 220 areas were polled. Polls in the home and neighborhood are part of this. Individual and family surveys frequently include questions about health, nutrition, health indicators, and health insurance.

Health and retirement research (HRS)

An aging population is a symbol of economic and social success as well as a source of new concerns. There is a significant social issue that must be addressed. Data warehouses that enable multidisciplinary research into the effects of an aging population on healthcare utilization are thriving. The use of traditional data collection methods complicates statistical analysis.

The University of Michigan's Health and Retirement Study (HRS), funded by the National Institute on Aging (grant number NIAU01AG009740) and the Social Security Administration, has been collecting data on retirees since 1992. Every two years, persons above the age of 50 are provided access to a series of in-depth interviews on a variety of themes. The HRS database can assist doctors and social scientists in better understanding the aging process. Canada, Mexico, the United Kingdom, Europe, South Korea, Japan, Ireland, China, Indonesia, Costa Rica, New Zealand, Brazil, and Africa have all conducted global aging studies.

Many similar cases can be discovered in the new HRS database. Both public and private HRS data are used in this study. To access private health information, a separate application is required, although anyone can sign up for the HRS data download portal to obtain public data. HRS includes research, RAND, economic, and cognitive activity statistics, as well as biannual data packages. Stata, SPSS, and SAS all have subdatasets available.

Dryad

Large dataset proliferation has reenergized worldwide efforts to standardize and distribute data. During the last decade, there have been significant changes in the architecture and protocols used for data management and exchange. All significant NIH contracts have been required to include a data disclosure study since 2003. PLOS One, the largest open-access journal, requires authors to submit information before publishing. BMJ Publishing Group encourages manuscripts containing Dryad data. Dryad is a platform that encourages data exchange and reuse.

Dryad is a non-profit membership organization founded in September 2008 and funded by the National Science Foundation. Dryad contains data in the domains of biology, ecology, and medicine. It is completely free to download and use. Dryad (http://dryad2.lib.ncsu.edu/pages/organization) was founded by top publications in biology and ecology, as well as scientific organizations, to archive and transmit data. Dryad provides free data security and reusability guarantees to researchers. The Dryad database housed 60,000 data sets and received 2,300,000 downloads in February 2018.

The findings of a study might be published in a variety of periodicals. Improved data sharing and reuse in medicine will result in new discoveries. Dryad stores data from scientists. Because of the results of Dryad searches, academic researchers and publishers can exert more editorial control over the publications they create. Dryad assigns globally unique DOIs to data packets for simple referencing. Dryad reads whatever you send it. Examine the file for viruses, copyright restrictions, important information, and the ability to open it. Driedad verifies the accuracy of the data. Data that is time-stamped, keyword-indexed, and cites other sources. Unless the data supplier objects, the report and data set will be made available online. Due to the dynamic nature of these details, Dryad will verify and update the article's title, abstract, author, and so on following approval or publication.

UK bio bank

Researchers obtained access to data from the UK Bio Bank, the world's largest biological sample database, on April 30, 2017. Between 2006 and 2010, the UK Biobank enrolled half a million adults aged 40 to 69 to collect information on their health, family medical history, and medications. The genomes and biochemistry of fifteen million blood, urine, and sputum samples were studied. The database can store long-term medical records. The database is fully accessible to academics. Major human diseases are examined in terms of lifestyle, heredity, and the environment.

In 2014, around 100,000 Britons donated MRI and X-ray images of their bones, organs, and brains to the UK Biobank. Computers are used to store medical imaging studies. This will be the largest imaging research of its kind. Cancer, heart disease, diabetes, arthritis, and Alzheimer's will all be reconsidered as a result of these massive data sets.

UK Biobank invites universities and other educational institutions to contribute their most recent academic results in order to assure the best quality of research.

Biologic specimen and data repositories information coord inating center (Bio LINCC)

Bio LINCC was founded in 2008 by the NHLBI, an international pioneer in the study of blood, lung, and heart illnesses.

A description of gene expression dynamics improves the quality of basic, applied, and clinical research. Bio LINCC contributes to the NHLBI's success by making scientific data and biological samples available to researchers. The Blood Disease Resources Department has maintained the NHLBI's biological sample bank since 1975, and the Cardiovascular Science Research Center has supervised it since 2000.

The Bio LINCC website went live in October of 2009. The portal receives clinical, epidemiological, and biological data from over 110 research groups. According to a 2015 poll of linked institutions done by the Yale School of Medicine, more than 90% of Bio LINCC users are satisfied, and their data may be used for clinical research. 73 percent of published authors have done it a thousand times or more.

Researchers can utilize Bio LINCC's data and samples for free, but they must bear their own transportation fees. Any requests for data or samples from the research community must be approved by the Bio LINCC. The NHLBI reviews all data and tissue sample requests that are submitted. The NHLBI determines research ethics after reviewing application material, study designs, and ethics committee deliberations. Bio LINCC's annual "paper submission reminder" day is March 1st. After getting accepted, researchers can update their application profile. The article is available on the study's official website.

GEPIA

Cancer genome research has benefited from large data sets. Hereditary cancer is caused by changes in cellular gene expression. Researchers will have access to additional sequencing data when more databases become available. A revolutionary web application called GEPIA (Gene Expression Profiling Interactive Analysis) profiles and assesses gene expression in cancer and normal tissues to fill in the gaps in cancer genomics data.

GEPIA was founded by Peking University Professor Zhang Zemin. GEPIA employs RNA-seq technology from UC Santa Cruz's Xena software. TCGA and GTEx data were used to examine RNA sequencing expression data from 8,587 controls and 9,736 tumor samples. These guidelines were developed using 9,736 tumor samples from 33 distinct types of cancer. When tumor and standard data are discordant, GEPIA leverages GTEx data to avoid costly identification procedures. GTEx sequenced almost 8,000 reference RNAs. To ensure consistency, the UCSC Xena project employed conventional methodologies to rebuild raw RNASeq data from TCGA and GTEx. The expression analysis offered by the TCGA and GTEx data is comprehensive. TCGA and GTEx expression data are generated in a comparable manner, allowing for direct comparisons. MySQL databases can be created with GEPIA. For topic analysis, R/PerL is employed. The GPIIA analysis, which is provided in an easy-to-use PHP interface, includes tumor/normal differential expression profiling and section localization based on tumor type or clinical stage. Modules are used to do analysis, gene comparison, patient survival, data correlation, data reduction, and the generation of personalized medicines.

The cancer genome atlas (TCGA)

Oncologists' traditional focus has been on tumor prognosis, early diagnosis, customized treatment, and prevention. Cancer may affect more than 20 million people by 2025, up from an estimated 14.1 million new cases in 2012. According to research, genetic variability is merely a small biological cause of cancer. That is why many cancer professionals are becoming interested in molecular genetics. The capacity to quantify gene expression enables for more precise cancer detection and treatment. The cancer genome is being studied using bioinformatics and whole-genome sequencing.

In 2006, the NCI got more government funding than the TCGA. Cancer research sponsored with $275 million in 2008 and 2009 is now available. In 2014, we added 33 new species to our list. Over 11,000 tumor samples with up to 255T of clinical, DNA, RNA, protein, and other multilayer cancer data are represented by ten uncommon malignancies. The information we gathered was beneficial. The Cancer Genome Atlas (TCGA) evaluates, identifies, uncovers, and characterizes all changes in human tumor genomes using high-throughput genome sequencing and gene chip technologies. Cancer researchers can use the TCGA's genomic and clinical data to better understand the illness and devise better prevention, diagnosis, and treatment options. The

Cancer Genome Atlas incorporates transcriptomic, epigenomic, and clinical information in addition to genomic and proteomic data. Some scientists combine gene expression and survival data to predict mortality. A wide range of groups monitor these materials.

The TCGA has enabled researchers to measure cancer growth at the cellular level, which has had a significant impact on tumor molecular biology and precision medicine. The TCGA data was used by the researchers to discover previously unknown mutations, intrinsic tumor classifications, and similarities and differences among distinct pancancers. It was discovered how cancers evolved. TCGA data will be improved by the use of bioinformatics technologies. Cancers include All-Cell Leukemia, Multiple Myeloma, Non-Hodgkin Lymphoma, and Osteosarcoma. The TARGET project uses chip and sequencing approaches to evaluate genomic and transcriptome data from children tumors. A molecular change map of each type of cancer can be constructed using multiomics. We can find cancer-related therapeutic targets and prognostic markers, as well as gene modifications that originate, enhance, and sustain the disease, by computing and validating biological processes. ALL and NBL launched the TARGET test. The five TARGET efforts are ALL, AML, KT, NBL, and OS.

Gene expression omnibus (GEO)

GEO, a global database of gene expression for public health, was produced by the National Center for Biotechnology Information (NCBI). Users and researchers now have several options for including, saving, and retrieving a wide range of data types. GEO's simple submission process makes advantage of researchers' data. All geodata contributions must use the MIAME format. The GEO database's structure makes it straightforward to download and query gene expression profiles for diverse disorders. The GEO database stores maps and raw data. GEO stores information in databases for platforms, samples, and series.

GEO dataset search results include the following information: name, description, species, platform, submitter contact information, series, publication date, numerical type, and sample count. The GEO expression map search can be used to view gene expression levels in each sample. Our research led us to experimental setups where we could see gene expression in a variety of conditions. Each dataset is accompanied by a report that describes the study, its data collection process, samples, and series, as well as their intended use.

Global burden of disease (GBD)

The public has historically been afraid of potentially lethal infections. In 1988, the Bill and Melinda Gates Foundation provided financing to the World Health Organization, the World Bank, and the Harvard School of Public Health to conduct a detailed evaluation of the world's disease burden. By shedding light on the cause and treatment of sickness, this has a positive impact on local health, social stability, and economic development. Health care is also getting better.

GBD has conducted extensive research on health deterioration. The database contains information on diseases, dangers, causes, injuries, and natural disaster-related symptoms. GBD can have devastating effects and even result in death at any age.

Costs are calculated using DALYs, incidence, morbidity, mortality, HALE, MMR, and exposure. Calculate your chances of death and other quantitative parameters. From 1990 to 2017, data on gender, age range, body mass index (BMI), and measurement type are available. Some of the study topics available on the World Health Organization's website include superregions, countries, and subnational GDP units. Many medical researchers take advantage of no-cost data sets.

3. CLINICAL DATA MINING METHODS

As medical technology advances, more data can be extracted. Because of technology improvements, medical records and follow-up data are now more accessible than ever. By uncovering patterns or laws in medical data, you can help patients with their prognosis, diagnosis, therapy, early detection, and cure rates. Data mining, as opposed to more traditional research methodologies, can expose facts in any given situation. New and relevant data is required. Statistics can be trusted more with data mining. Data mining can be used to both describe and prescribe a path of action. The importance of data integration and cluster analysis is emphasized. For forecasting, use classification and regression analysis on existing data.

Description Association analysis

Association analysis, also known as association mining, is the activity of examining multiple forms of data (such as financial, social, or geographical data) in order to uncover potential relationships between distinct behaviors or objects. Patterns in large datasets can be discovered via correlation analysis. The well-known "shopping basket" experiment demonstrates correlation. This is accomplished by analyzing clients' purchase habits depending on what they place in their virtual shopping carts. Shops can use customer purchase data for marketing purposes. The ability to sell in a retail situation. Association analysis compiles the most common items and the rules that govern their occurrence. The second priority should be rulemaking. The rule is an association rule if the high-frequency item group from the first phase meets the confidence criteria. In association analyses, machine learning techniques such as FP tree frequency set, Upgrade Lift, and Apriori are applied.

Apriori algorithm

All nonempty subsets of sets of often occurring items are frequent, while all supersets of sets of rarely occurring items are uncommon, according to the a priori technique. I rarely visit a website that lacks a few critical things that I demand. The most frequently purchased items can be seen in the purchase records. In chaotic data with a predictable structure, there is a "pattern" as well as low- and high-frequency modes. Most people would prefer more frequent classes. Apriori can be used to find collections of frequently recurring objects to discover the "frequent mode," a high-frequency mode. Apriori reduces the number of queries on item sets and increases their speed.

FP tree frequency set algorithm

The FP tree is constructed over time by sequentially reading transactions and allocating them to nodes in the FP tree. When two transactions have comparable portions, they can share a route. The FP tree compression improves when numerous paths converge. A compact FP tree in memory can obtain groupings of frequently requested objects instead of scanning and storing data on a hard disk. The data is retained, and the condition bases are mined independently using the FP tree frequency set technique, which condenses the frequency set after the first passthrough into a frequent pattern tree.

Upgrade lift

Inadequate data does not ensure accurate Apriori or FP tree frequency set rules. Lift's evaluation considers quality characteristics. Lift indicates the strength of a randomly selected predecessor and back piece by increasing the possibility of the next front piecerithm piece. Even if the rules are followed and everyone understands them, mistakes might happen. Lift is an innovative metric for determining the effectiveness of evaluation criteria. As the risk of the previous front piece occurring at random is represented, the likelihood of the subsequent front piece grows.

Cluster analysis

Each dataset must be appropriately classified by the classification method. Cluster analysis is required if the aforementioned conditions are not met. Cluster analysis is used to group things with similar characteristics. Clustering is a statistical approach for grouping data items with similar characteristics. Clustering methods span from grid and splitting to hierarchical and density-based.

Partition-based algorithm

K-means is the most widely used cluster analysis approach. The benefits of both the prototype and partitioned distance techniques are obvious. We utilize K to classify N items, and then we employ a perfect idea to correct any errors. The K-means technique is simple to apply, productive, and reliable. In high-dimensional data, poorly implemented K-means cannot find non-spherical clusters.

Hierarchical clustering algorithm

Within the aggregated information, hierarchies exist. Clusters can be further subdivided or combined using descending hierarchies. Hierarchical clustering algorithms such as BIRCH, CURE, ROCK, and Chameleon are widely employed. The method starts with a point combination. Structures inside a cluster can have an impact on how closely they are grouped. The merging process is complete when further permutations produce solutions that are unsatisfactory for various reasons.

Density-based algorithm

The density approach divides the data space into dense and sparse regions to discover clusters of various sizes and shapes. The most well-known approaches are DENCLUE, OPTICS, and DBSCAN. DBSCAN scans are commonplace. It groups together local problems and needs into interconnected clusters. Primarily to eliminate background noise. The density of O varies based on the number of items in its vicinity. The project is seeking for locations that are both central and out of the ordinary. DBSCAN's cluster classifications are based solely on geometric information. Due to its temporal complexity, the algorithm is incapable of processing high-dimensional data.

Grid-based algorithm

An index of non-convex clusters cannot be constructed using partitioning or hierarchical clustering. Even the slowest density algorithms can identify concentration. Between 1996 and 2000, data miners developed grid-based grouping techniques.

A density-sensitive grid simplifies the processing of algorithms. Multiple grid resolutions are utilized in grid-based clustering. This approach requires a certain number of elements in each dimension of the quantization space. Among the well-known techniques are WavemCluster, CLIQUE, and STING. STING employs grid multiresolution to partition space into square units of varying resolutions for high-dimensional subspace clustering, whereas WaveCluster and CLIQUE depend on wavelet analysis and grid and density clustering, respectively. Cluster shapes can be discerned.

Prediction

Regression analysis

Linear regression analysis requires two independent variables. Use on a regular basis. Y = w'x + e when the normal distribution has no mean. A linear regression or a multiple linear regression can be performed depending on the number of variables being studied. A straight line is a rough approximation for a single independent and dependent variable in linear regression. The hypothesis of a linear relationship between the explanatory variables and the dependent variables is the foundation of multiple linear regression.

Several elements frequently interact to produce an impact. A regression requires at least two independent variables. Multiple-parameter regression analysis. Predictions based on a high number of independent variables are more likely to be realistic and correct. As a result, using multiple linear regression rather than just one is preferred. Finding the partial regression coefficient for each independent variable, as well as generating a regression equation and testing the null hypothesis, is at the heart of a multiple linear regression study. Remove the partial regression coefficients for insignificant variables and recalculate the multiple regression equations. Using the least squares linear model.

Classification analysis

Managers research information categorization. Use tags to rapidly find material that fits your needs. Accurate classification can improve the outcomes. Profit regressions, logarithms of returns, and machine learning are all well-known methodologies. Classification models improve data comprehension. Expenses may be restricted. All data sets require categories and antecedents. When working with a categorical dependent variable or a large number of independent factors, the classical statistic is inapplicable. Machine learning has made sophisticated data analysis considerably more precise and practical.

4. PROSPECTS AND CHALLENGES OF MEDICAL DATA MINING

New enterprises are connecting precision medicine and traditional treatment by collecting and analyzing massive volumes of data. Big data's full promise for global precision medicine and new health management has yet to be completely realized. Forecasting the future of big data analysis, data visualization, and artificial intelligence is achievable with well-planned investments in platforms that promote technological and human development. Large datasets rarely reveal unexpected trends. Prepare for some life-changing, medically beneficial changes. Big data provides numerous options. Medical big data mining is hampered by high data modality, latitude, type, and structure, as well as the complexity of medical knowledge ideas and the lack of technical advances in medical knowledge reasoning. Neither the outpatient procedure nor the hospital's electronic medical record system are regulated. While analyzing life and health data on a broad scale is difficult, advancements in infrastructure, technology, and human resources may make this work more viable.

5. CONCLUSIONS

Data mining techniques and large-scale database systems are discussed in this article. As medical technology advances, more data can be extracted. Because of technology improvements, medical records and follow-up data are now more accessible than ever. Data analysis based on trends or correlations could assist medical care. Encourage early diagnosis, a good prognosis, and effective treatment. COSMIC, HGMD, Oncomine, cBioPortal for Cancer Genomics, SRA, the WHO Mortality Database, Orphanet, DGV, and OMIM are all good places to start. Medical data mining will have an impact on hospital diagnosis, treatment, research, education, and administration as its theory and practice evolve.

REFERENCES

1. Schlick CJR, Castle JP, Bentrem DJ. Utilizing big data in cancer care. Surg Oncol Clin N Am. 2018;27:641–652.
2. Trifiro G, ultana J, Bate A. From big data to smart data for phar- macovigilance: the role of healthcare databases and other emerging sources. Drug Saf. 2018;41:143–149.
3. Binder H, Blettner M. Big data in medical

science–a biostatistical view. Dtsch Arztebl Int. 2015;112:137–142.

4. Bahi M, Walmsley RS, Gray AR, et al. The risk of non-melanoma skin cancer in New Zealand in inflammatory bowel disease patients treated with thiopurines. J Gastroenterol Hepatol. 2018;33:1047–1052.
5. Jonathan E, Mayer RHG. Arsenic and skin cancer in the USA: the cur- rent evidence regarding arsenic-contaminated drinking water. J Der- matol. 2016;55:585–591.
6. Bayne LE. Big data in neonatal health care: big reach, big reward? Crit Care Nurs Clin North Am. 2018;30:481–497.
7. Ristevski B, Chen M. Big data analytics in medicine and healthcare. J Integr Bioinform. 2018;15: 20170030.
8. Bellazzi R. Big data and biomedical informatics: a challenging opportu- nity. Yearb Med Inform. 2014;9:8–13.
9. Sinha A, Hripcsak G, Markatou M. Large datasets in biomedicine: a dis- cussion of salient analytic issues. J Am Med Inform Assoc. 2009;16:759– 767.
10. Scruggs SB, Watson K, Su AI, et al. Harnessing the heart of big data. Circ Res. 2015;116:1115–1119.