

Big Data Solutions For Mapping Genetic Markers Associated With Lifestyle Diseases

Kiran Kumar Maguluri¹, Zakera Yasmeen², Rama Chandra Rao Nampalli³

Abstract

Lifestyle diseases are present in multitudes and strike indiscriminately, leading to chronic illnesses. There are no precise answers to many of the questions that perplex the general populace. Identifying individuals who need additional surveillance to prevent the onset of these diseases is increasingly becoming a major priority. The big data era and the advent of high-throughput technologies have led to a significant reduction in the cost of genotyping thousands of individuals to identify genetic markers associated with complex diseases. Searching for tiny genetic signals amidst millions of DNA variants means analyzing tediously large data sets. We discuss big data analytics as¹ a solution for the identification of genetic markers associated with lifestyle diseases. Given the limitations of predictive modeling of complex diseases despite having powerful predictive tools, we propose a new matrix-based gene-gene variant association test called Mod-Log. Using this approach, we identify thousands of genetic markers that are predictive of health-related traits in a large data set of 15,000 aging men, showing a strong association between gene-gene interactions and complex diseases.

Keywords: Lifestyle Diseases, Chronic Illnesses, Surveillance, Big Data, High-Throughput Technologies, Genotyping, Genetic Markers, Complex Diseases, DNA Variants, Data Analysis, Genetic Signals, Big Data Analytics, Predictive Modeling, Matrix-Based Test, Gene-Gene Variant Association, Mod-Log, Health-Related Traits, Aging Men, Gene-Gene Interactions, Predictive Tools, Genetic Studies.

1. Introduction

The focus of big data initiatives related to human health was initially on sequencing individual genomes to find genetic variants predisposing to disease. A more comprehensive understanding of disease mechanisms, emerging novel treatment options, and recognizing interactions of genetic susceptibility variations with lifestyle choices and environmental exposure requires combining single nucleotide variations and structural variations from entire genome sequencing with many other data sources. For most conditions of interest, a thorough investigation of big data resources exposes a continuum of sophistication, ranging from specific studies focusing on a particular question to comprehensive national big data resources of extensive scale and scope. Several countries have made substantial investments in health big data for general health surveillance to maintain a comprehensive overview of citizens' health and well-being. The main focus of health surveillance is on information in the public health

¹IT systems Architect, Cigna Plano Texas, kirankumar.maguluri.hcare@gmail.com, ORCID: 0009-0006-9371-058X

²Data engineering lead Microsoft, zakera.yasmeen.ms@gmail.com, ORCID: 0009-0004-8130-2111

³Solution Architect, Denver RTD, Parker, CO-80134, nampalli.ramachandrarao.erp@gmail.com, ORCID : 0009-0009-5849-4676

domain, from a disease perspective. Investing in extremely large-scale resources for various health issues will identify key questions synthesizing an array of studies for in-depth investigation. A key aim is to provide value for money and high data quality, by recognizing that the best methods for data handling and analysis are data-driven and require input from a wide array of core competencies. Interoperability is key for the survival of systems aiming to solve broad health information objectives related to monitoring and analyzing citizens' health and well-being. Awareness and implementation of advanced analytical techniques and methods to enable desired outcomes from integrated combined health data should be facilitated by government involvement.



Fig 1 : Genomics and Big Data Analytics for Personalized Medicine and Health Care

1.1. Background and Significance

Genetic markers can be effectively mapped using the DNA sequences of populations from multiple countries that have been made publicly available. However, most of the currently used marker mapping systems do not provide the associated weights of genetic marker groups that are needed as inputs for the creation of classification models that can detect lifestyle diseases. In this paper, four extraction systems are compared based on two classes of widely used model-free variable weight extraction methods for establishing genetic marker groups associated with lifestyle diseases. To make the results more relevant, a primary study is conducted separately for 101 Asians, 131 Africans, 345 Europeans, 204 Mexicans, and 463 Americans of African ancestry using 71,888 DNA data. The results for people of the same ethnic group are compared, and a sensitivity analysis is conducted to determine the best extraction method that could guide further research to minimize the high cumulative percentage of lifestyle diseases.

It is well known that the environment and the genes of humans can affect whether or not a person can have lifestyle diseases, such as diabetes, coronary heart disease, obesity, hypertension, and certain cancers and cardiovascular diseases. The genes that are involved in regulating vital biological pathways and their interactions have been identified. However, even using big data, the effects of multiple gene-gene interactions and the effects of the genetic markers of different populations on lifestyle disease predictions are unclear. The mission of this paper is to solve such problems so future precision health scientists can understand the genetic factors that predict lifestyle diseases and develop diagnostic methodologies that can be translated to different populations.

1.2. Research Aim and Objectives

The research aims to provide an overview of current big data solutions for mapping genetic markers associated with lifestyle diseases. The objectives of the research include developing an analysis and comparison of the main statistical and software tools used to search for genetic markers; theoretical refinement of the developed methods, and the formation of approaches to address the identified disadvantages of software tools. Review of public software and packages; analysis and development of a classifier based on publicly accessible programs and data for

various problems of lifestyle disease prediction. During the research, special emphasis will be given to statistical methods and software packages related to solving selected analysis problems using ready-made tools, such as SNP, histograms, or other big data sets. To implement analysis tasks, program modules based on a set of tools and scripts providing integration of all prepared parts will be developed. After the simulation, an efficient software package will be implemented.

Such a study is important due to its practical significance, as it will be possible to apply the received recommendations to real problems of predicting lifestyle diseases and developing treatment methods. The developed software will be tested on available data, and the obtained results will be analyzed. The implementation base of the project includes state-of-the-art statistical software such as SNP, histograms, and a set of statistical tests for SNP associations. It is expected that the implementation of the project will allow for the development of a comprehensive and automated computational solution aimed at predicting the results of lifestyle diseases that can be used by a wide range of researchers engaged in related issues. The developed software package should be freely distributed and installed at genome analysis centers.

2. Understanding Lifestyle Diseases

Lifestyle diseases are defined as those that are linked to a person's lifestyle. They include heart disease, obesity, and cancer, among others. Increases in population and wealth contribute to the growing demand for food products such as meat. This increase, combined with recent productivity improvements, may compromise dietary guidelines and, consequently, human health. A source of this challenge arises from the tension between breeding and genetic matching of livestock for production versus choosing breeds of animals and animal products that have attributes characteristic of a healthy diet. On the other hand, many patients with lifestyle diseases (or their families) request guidance on what to eat or on what symptoms can be relieved through eating a healthier diet. Scientists have developed methods for genetic matching of meat attributes to human dietary guidelines.

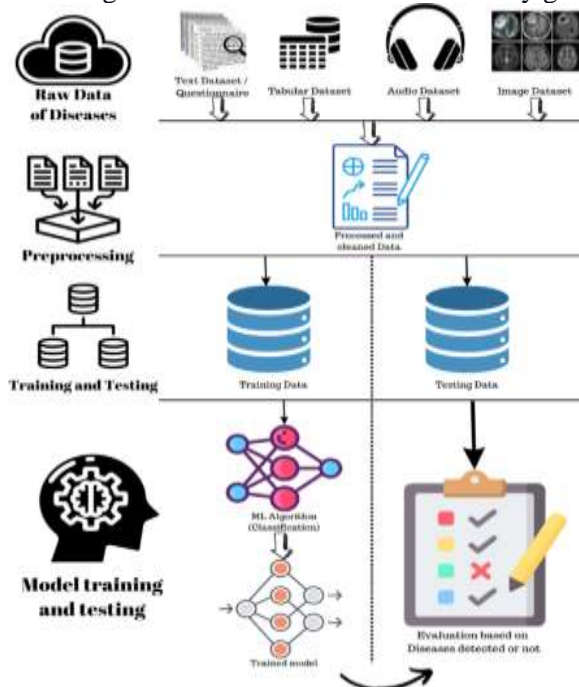


Fig 2 : Prediction and Prognosis of Lifestyle Diseases

Fatty acids are concentrated in milk and meat. In general, consumption of foods rich in essential fatty acids has beneficial effects on human health. However, in specific situations, some fatty acids have been linked to negative effects, such as the increase in LDL cholesterol in the blood, which is associated with an increased risk of developing heart disease. In this context, many studies aim to quantify genetic influences on milk fatty acid content to control the increasing prevalence of dairy cattle rich in harmful fatty acids. How the genetic influence of a breed affects meat or milk content with favorable fatty acids and thus benefits human health supports a dual objective: to breed informed livestock while providing a health-oriented dietary guide – laboratory and nutritional data at the same time. By continuously identifying DNA blocks involved in final phenotypes composed of high genetic values, quantitative trait locus methods increase the success of predicting complex traits for genomic selection while determining the genetic basis of traits. Fast and efficient detection of trait-associated DNA fragments calls for comprehensive sequence resources linking DNA markers to genomic positions and providing unique sequence information within the incorporated species.

$$S_g = \sum_{i=1}^n w_i x_i + b$$

Equation 1: Genetic Marker Association Score

S_g : Genetic marker association score.

x_i : Genetic feature or variant i .

w_i : Weight for genetic feature i .

b : Bias term.

2.1. Definition and Types

Human diseases are caused by a variety of reasons. Many are directly related to the lifestyle of the individual. All lifestyle diseases have underlying genetic factors, which make certain individuals more susceptible to the diseases. Identifying these genetic markers would be the key to developing genetic tests to assess the risk for lifestyle diseases. Identifying the genetic markers could help us to be proactive in preventing children at risk from adopting lifelong damaging habits that one or both parents possess. It could also be a way to fine-tune the supplements, diet, and lifestyle habits by our genetic risk. The purpose of the study is to illustrate a big data problem, to test the usage of leading processor technology, and to make it cheaper and faster for researchers in biotechnology to extract more information about the human genome.

The explosion in the amount of discovered data will need large-scale data analytics platforms to discover insights and find associations that have not been found. To use the potentially increased knowledge in the biotech industry, it is important to be able to make sense of the astonishing amounts of parallel data. This study will explore mapping markers with significant influence on four lifestyle diseases – cancer, diabetes, heart diseases, and obesity – and use them as examples to show how the processing could be done more efficiently by exploiting technologies. The rate at which massive datasets are generated for genomics research motivates the exploration of scalable and performant solutions for genetic marker set analysis. We discuss the typical exploratory nature of genetic marker set studies that test a variety of derived genetic marker associations in search of phenotype-dependent effects.

2.2. Epidemiology and Impact

Epidemiology of Type 2 Diabetes

The epidemic of diabetes has increased disturbingly in recent years, both in situations of pre-diabetes stages, and subsequently manifested the horrible epidemic in deficiency of insulin secretion accompanied by impairment of glucose metabolism. There are up to 400 million diabetics worldwide in 2014, and this number is projected to double by 2040. Asian ethnic groups have one of the highest diabetes susceptibilities, and it is projected to take a large proportion of the global diabetes burden. With the surge of the aging population and the increased impact of gestational diabetes and prediabetes, it is necessary to propagate the lessons from the past and provide a more aggressive strategy for diabetes prevention in Asian communities. The Asia-Pacific region comprises several countries with profound ethnic diversities, so implementing effective means of prevention or intervention is challenging. However, more activities in lifestyle intervention, like nutrition-related studies for high-risk groups, should be adopted before insulin resistance establishes type 2 diabetes.

Epidemiology of Hyperglycemia

More than a quarter of Chinese adults have glucose levels qualifying them to be diabetic. However, even at fasting plasma glucose slightly above 95 mg per deciliter, the adverse and expanding spectrum of blood pressure and cardiovascular disease that has been widely discussed for increasing fasting glucose is not attenuated by the therapeutic measures used to reduce fasting glucose levels to within the conventional normoglycemic range, including diet and various medications. Therefore, it is suggested that fasting plasma glucose concentrations be lowered from the highest end of the accepted normoglycemic range.

3. Genetic Markers and their Role in Lifestyle Diseases

Lifestyle diseases are a class of non-communicable diseases that are associated with the way a person or a group of people live. They are usually developed as a result of their lifestyle, which is their characteristic behavior patterns, associated with the way they can afford to live – for example, what they eat, their physical activity levels, what they drink, whether they smoke, and their satisfaction in their working environment. Lifestyle diseases are diseases that appear in higher numbers as countries become more industrialized and people live longer. They are thought to be developing as a result of an increasing sensitivity of genes to the environment, particularly to changes in diet. The contributions of human genetic variation underlying these disease conditions are not as simple as originally thought. There are no simple relationships between disease and genes. Genes are not even necessary or sufficient for a condition to occur. The relationship between genes and their effect is more complex.

Genome-wide association studies focus on relatively common genetic variants that are expected to have small effects, whereas rare variants are expected to have larger impacts on common diseases. These studies are predicated on the assumption that genetic variation in a sufficient number of cases and controls will be found to be associated with specific diseases and that the underlying genetic effects are approximately equivalent among different backgrounds. Today, the relationship between genetics and lifestyle diseases is not well understood, with much information coming from these studies. There is not a complete understanding of all the polymorphisms that can influence lifestyle diseases, and of those that are known to have an influence, much of the relationship to the lifestyle disease is not understood. There are also environmental and epigenetic processes to consider.



Fig 3 : Implications in chronic disease

3.1. Overview of Genetic Markers

In this section, we would like to introduce several big data-driven solutions available for mapping genetic markers associated with lifestyle diseases. We will start with a brief introduction to genetic markers and the study of genetic markers and lifestyle diseases. Genes are the building blocks of heredity. They are passed from parents to children and contain the instructions for building and maintaining the body's cells, tissues, and organs. Unlike lifestyle diseases that result from complex interactions among lifestyle behaviors and environmental exposures, conditions, and genetic variants, lifestyle disease risk assessment has been considered as the individual's lifetime risk. Major research has been conducted to study markers for lifestyle disease prediction.

Given the enormous amount of data readily available, easy and efficient data access is required for building lifestyle disease risk prediction models. Big data technologies provide high-level solutions to challenges in accessing and analyzing the data. A large number of known genetic markers are available for public use. An elastic data warehouse can handle many users and terabytes of data and has been proven to support quick response times. This project lifecycle has the characteristics of big data and requires collaborative problem-solving through the use of advanced runtime infrastructures.

3.2. Association with Lifestyle Diseases

Big data solutions facilitate the determination of potential arrays whose marker distribution can be associated with lifestyle diseases. Recent work within the big data domain has been designed to provide results at a much lower cost and possibly lower the gold standard for genetic variation assessment. Most of this work has taken on the dimension of a case study and has focused on generating results in response to a specific scientific question or challenge. Given the outstanding technological progress and the relatively crude phenotypic measures employed in some of the published work, carefully designed and significantly larger scale projects can provide important clues as to how to interrogate the complex issue of genetic risk scores calculated from a wide range of validated genetic variants.

Current big data bioinformatics and genome-wide technology advances provide fast and cost-efficient generation of a large amount of variant data at an individual level. One promising approach to pinpoint new risk factors is to leverage and integrate this amount of data to identify

genetic variation by association with expression profiles of disease-relevant tissues. These methods include the global biologic assay combined with next-generation sequencing solutions, allele-specific silencing, mapping, and RNA sequencing. Integration of SNP genotype with mRNA transcript abundance level yields gene expression quantitative trait loci, which represent genetic polymorphisms regulating mRNA expression of target genes. In this review, we introduce a wide range of methods captured by the functional genome-wide association study and the translational genomics for the potential to translate arrays generating data for applied findings that will further improve public health.

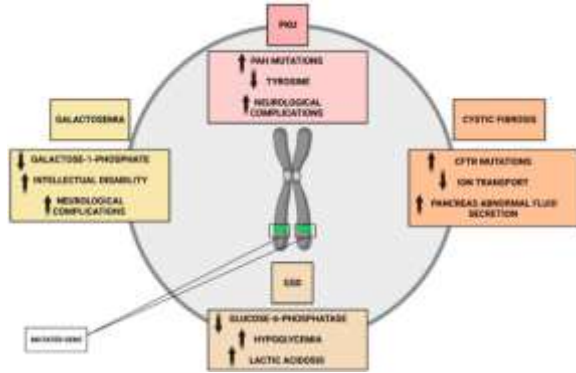


Fig 4 : Potential Therapeutic Interventions in Metabolic Diseases

4. Big Data in Genomics Research

The cost-cutting of next-gen sequencing, as well as the even faster data generation, are pushing molecular medicine into the big data era. It is suggested that in 2020, the data volume of genomic information archived on tape will reach 2.6 zettabytes. This might be an underestimate. The fact is, over the past 14 years, between 2001 and 2015, whole-genome sequencing alone generated 20 zettabytes of raw data. For the study of diabetes, one popular big data project is the T1DGC-T1D Genome and Microbiome projects for the primary prevention of type 1 diabetes. Four major projects under the umbrella use various sequencing techniques to collect genetic data on thousands of diabetic patients and control subjects. In collaboration with industry, the researchers are also planning machine learning for feature extraction from the microbiome datasets.

One solution to efficiently manage big data from human genomics research is a distributed infrastructure. The next-generation genome analyzer is designed to be used as a grid, composed of a cluster of nodes dynamically balanced by the data manager. Other similar horizontally scalable platforms with elastic storage include next-generation sequence analysis platforms and commercial clouds with an on-demand fee. They can all manage petabyte-size data storage. Even among many computing platforms, there exist several open-source and commercial software solutions for the quality management of big genomics data obtained from the next-gen sequencer. Genomics researchers can compare the QC tool of their choice for specific diseases.

$$I_{gg} = \prod_{i=1}^m x_i^{\alpha_i}$$

Equation 2: Gene-Gene Interaction Prediction

I_{gg} : Gene-gene interaction score.

x_i : Genetic variant i from gene i .

α_i : Interaction coefficient for variant i .

m : Number of genes involved in the interaction.

4.1. Definition and Scope

Lifestyle diseases, usually a consequence of an unhealthy lifestyle, are a passive consequence of a person's way of life and include type II diabetes, obesity, atherosclerosis, ischemic heart disease, and hypertension, among others. These conditions trigger the onset of several other serious and persistent illnesses, such as chronic kidney disease, retinopathy, and neuropathy. The detailed study of the genetic markers linked to lifestyle diseases may improve the identification of at-risk individuals, contribute to personalized therapy, offer clear insights into the etiology of the disease, evaluate the environmental and genetic contributions throughout a lifetime, from pregnancy through prenatal to childhood and adulthood years, and facilitate the design of clinical studies, including clinical trials. Furthermore, genetics-based approaches offer the potential to modify the negative outcomes that result from adopting an inappropriate lifestyle and to contain healthcare costs, mainly those negative outcomes predicted to escalate rapidly.

The field that studies the connection between genes and lifestyle has its roots in genetics and the correlation of genetic content with specific behaviors. According to a widely adopted approach, a phenotype can be modified in response to a specific form of environmental stimulus during the lifetime of the individual. The association between a specific gene marker and the mode of life modification is the subject of lifestyle genetics, whose goal is the dissection of the genetic, epigenetic, transcriptomic, and metabolomic contributions to phenotypic adaptation. In the last 25 years, personalized medicine-oriented research has catalyzed substantial advancements in biotechnology and bioinformatics, leading to the emergence of big data in genetics or genomics, the study of the complete set of genes within and related to the organism or system holistically or globally.

4.2. Importance in Genetic Marker Mapping

One of the main goals of current genetics and biometry is to map the genetic markers that are related to susceptibility to the most common complex diseases. Discovering such regions might improve how we can fight them. One of the current preferences is to use the association of genetic markers with phenotype values, that is, to analyze how variations in the genetic markers influence the phenotype values. The set of variations of a DNA sequence among individuals is called a polymorphism, and can be a variation in the sequence that may alter the proteins, by one or more base pairs inserted, deleted, or replaced; or it may be a silent polymorphism that does not alter proteins but may change the way that the region is recognized and/or transcribed. Knowing the locations and the values of these polymorphisms is an important stage before the effect of polymorphisms on the trait is studied.

The large number of genetic markers analyzed now leads to what is called the problem with multiple tests, or a large number of comparisons, in which the hypothesis tests for all the genotypic effects for each polymorphism must be done to correct the type I error. Due to the high level of association between these tests, p-value adjustments based on the number of comparisons must reach a very low p-value to detect a significant result. It is possible that markers with a moderate contribution to the trait are not detected because of this. It is common to observe many markers strongly associated, something that would be expected only from spurious associations. A set of methods to extract important information by analyzing large

The overall evaluation of the performance of models results from tests of observed and expected relationships. Previous conclusions, taken from cross-validation and inference on independent assessments, are based on cohort or trait distributions in the training cohort. Therefore, one major source of evaluation for model development comes from the interpretation of gene expression, biomarkers, and the application of findings. The extraction of knowledge to generate insights that can further explain the functioning of the genotype-to-phenotype relationship is a task expected from the interpretation of machine learning models built with human genomics data. These last steps are inherent to every analytical path in genomics medicine and can be taken from methodological proposals produced with biases in the development of predictions.

5.2. Data Mining Techniques

Data mining is a subfield of computer science that tries to find interesting patterns from large and complex databases. It is often described as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in the data. In the process of solving complex computational problems, if there is big data, then parallel computing is to be used, which can be classified into two types: the problem parallel model and the data parallel model. Because of the use of distributed architecture in big data, the data parallel model is to be implemented, and in the data parallel model, several simple parallel operators are to be applied. The parallel operators used in the data set are called data mining primitives.

Data Mining Primitives: The primitive set for data mining includes support, join, and prune or filter. These primitives are useful for almost all types of data mining algorithms: support, which is to filter all the transactions having frequent itemsets; join, which is to join the frequent itemsets of n items to find frequent itemsets of $n + 1$ items; and prune or filter redundant itemsets, which is to filter the infrequent itemsets. These primitives can also be used to implement various data mining algorithms. Data mining is applied to the output of large and high-dimensional data from various data sets such as gene expressions, SNPs, gene regulatory pathways, clinical data, tumor pathology, and statistical data. Data mining techniques may group genes based on their expression-function relationships, disease progression, or sample types. The application field of data mining algorithms also varies from microarray data to clinical outcome data. The steps required to apply data mining are as follows.

Data Cleaning: It is to improve the quality of data by preprocessing.

Attribute Selection: It is to neglect unimportant attributes from the data set.

Clustering: It is to group similar objects, thus knowing more about the data set.

Association Rule Mining: This is to identify the frequent patterns and mine the relationships in the data set.

Classification: It is to classify the data sets to assigned models or existing data sets.

Regression: It is mainly used to build the outcomes such as clinical trials for future predictions.

Prediction: For example, decision tree pruning involves removing the terminal nodes of the tree. If the relative proportion of error reduction, the relative error of the two child nodes plus the parent node is less than the delta, then the prune will be conducted.

Outlier Detection: When the outliers are labeled as anomalies or noise in the data set, they should be detected and possibly removed.

Translation: It is to transform a functionally related or associated group of genes to show its enriched biological significance.

5.3. Cloud Computing

The availability of numerous cloud-based solutions has made data storage, computing capacity, and software capabilities easily accessible. Setting up an on-premises computational environment is being gradually replaced by using rented computational resources as and when required. Various cloud service providers offer different forms of platforms as a service,

software as a service, and infrastructure as a service model, thus providing a service to meet the end user's computational and software needs.

With an increased focus on big data analytics, healthcare research projects are making use of public, private, and consortium cloud-based solutions. Cloud computing provides researchers with various systems as well as inexperienced programmers with the infrastructure required, provided they are given some kind of efficient mechanism to distribute their jobs and their data in one of the many forms of validation protocol testing. However, the industry is very much in a state of rapid development, and the full potential of this utility has yet to be realized, especially regarding the increase in storage requirements and associated costs. They also address an overlapping tool set; therefore, it is not clear to all the users which solution is the best one for their needs.

6. Case Studies and Applications

Genetic variations, mainly single nucleotide polymorphisms (SNPs), constitute the majority of human genetic markers. Their role in human traits and diseases is of high interest. SNP data is made available through various efforts that present millions of SNP locations in the human genome. The role of big data infrastructure in managing and analyzing such large datasets has allowed the development of efficient tools that can handle genetic data, including data management tools, imputation methods, association analysis, visualization, and query systems. Moreover, big data allows combining genotypic and phenotypic data for effective research.

The 1000 Genomes data, for example, comes as phased haplotypes at most SNPs. Its usage provides researchers with new information and opportunities for large-scale association analysis. Many of the common variants were imputed and made available through imputation services. Since the application of statistical association testing for a genome-wide set of biological markers is computationally expensive, a single test can require several minutes or even hours. Such tests have become feasible only recently, facilitated by powerful computational infrastructures. With the drop in genotyping costs, millions of genetic markers per individual can currently be analyzed. Such studies typically require analysis that takes advantage of the big data character of genetic data, whether they are piggybacked on other public datasets or intrinsic to the dataset, such as in gene expression or GWAS data. In this section, we describe a few data repositories and possible analysis pipelines available to researchers in genomics interested in detecting and utilizing genetic variation.

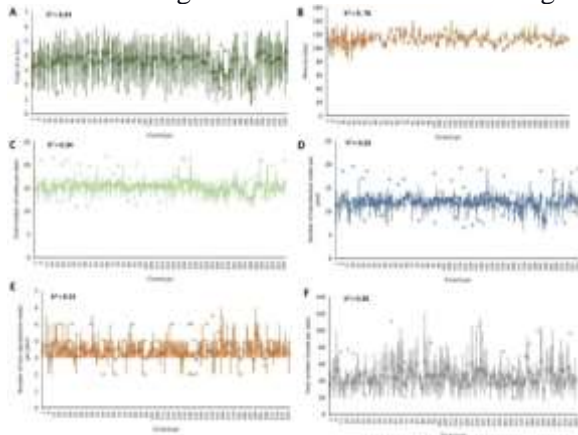


Fig 6 : Machine-Learning-Based Genome-Wide Association Studies

6.1. Real-world Examples of Big Data in Genetic Marker Mapping

More detailed big data applications for mapping lifestyle disease genetic markers are provided by real-life examples in research. The examples were based on a platform that details over 70,000 human genomes, all described by over 200,000 SNPs. This translates to over 14

petabytes of data that need real-time manipulations, such as reduction via PCA, large-scale data extraction for specific categories of subjects, machine learning for disease mapping, and association with clinical markers. Out of over 70,000 subjects, about 34,000 had expressed their consent to use the genotype and phenotype and took part in the first analysis. A Data Controller Request system was designed and implemented. A variant of the Generalized Pearson Association algorithm was implemented, and examples were applied in identifying associations between SNPs and the Body Mass Index, coronary heart disease, T2DM, cancer, internal vestibular implants, and age of disease onset.

15,000 SNPs from Chromosome 1 have been extracted from the larger dataset and were further used to train an sGBDT model as the general BMI predictor for subsequent personal genotyping. The data preparation time and the model training time were measured at several minutes and were less than 1 hour, respectively. The model can estimate if the correlation associations are stronger or weaker than the PCA non-correlated association. The binned mean for each BMI class can further be used to find the SNP frequencies for each BMI class in the training set and deduce the BMI intervals for a subject with an unknown genotype. Our model can predict BMI, CHD, and T2DM with real-world speed and accuracy, using a very small amount of genomic data, and is compliant with the next-generation healthcare concept of a validated personal service prototype.

7. Challenges and Future Directions

As an interdisciplinary research work, we needed to deal with several challenges mostly related to data analysis to identify reliable genetic markers associated with lifestyle traits using a high number of living beings under study. In this way, the most challenging aspect is related to the high number of phenotypic measures available for the HRS and f:m diagnosis. Since these high-dimensional feature sets can be very noisy, we adopted feature selection in all data analyses. We understand that this is beneficial to the performance of classifiers and to evaluate the impact of different cardinality feature sets in the way we use the selected genetic markers to accomplish our specific goals. However, issues related to class imbalance inherited in living beings also influenced the biomarker selection.

Since we are still analyzing the full collection of the phenotypic measures, such as the family of group classes we created from a random and equal split of the f:m class and using the group time of minimum temperature responses, we have several other challenging issues related to the fact that some of the considered features are known to be highly correlated and have minimal clinical effects. To understand and see the impact of these analyses aimed at identifying such predictors, we made use of several predictive models avoiding arbitrary splits in our feature sets to embrace the temporal domain, such as well-known single and ensemble machine learners like Random Forest, Naive Bayes, kNN, SVM, and Artificial Neural Networks.

7.1. Ethical Considerations

The deployment of big data technologies in healthcare, coupled with the establishment of huge biobanks, has promoted a wave of analysis to identify, examine, and track genetic markers. The advances come with a spike in data-driven research on neurological disorder subtyping, both for autism spectrum disorder and Alzheimer's dementia. Additionally, research is migrating as the costs of data collection from fewer genes have given way to collection on full-genome sequencing. The dynamics can lead to associations between more genes with lower marginal impacts or the same with neurobehavioral traits. However, these toolkit generalizations do demand that specific tinkering be updated and bespoke for the particularities of these diseases. Their biological relevance is now more pertinent than ever before. The use of neuroimaging genetics in particular opens up a territory that requires novel ethical considerations for managing, analyzing, and distributing results. These considerations are focused on the realms

of data privacy, the sensitivity of results, and informational risk. They are substantially more pronounced when applied to neurological health data due mostly to the increased scale of potential harm and the increased nature of selectivity. The research further probes mental health inequalities and the stigmatization of potentially harmful genetic information that could be calculated and used to discriminate. These big data-driven analytics could uncover a final layer of ethical issues related to machine learning and the optimization of clinical decision-making. Its algorithmic biases raise doubts of an operational nature, and it could also deliver the chancy implications of rectification by neuronal pathway manipulation, rekindling memories of unsavory control feelings.

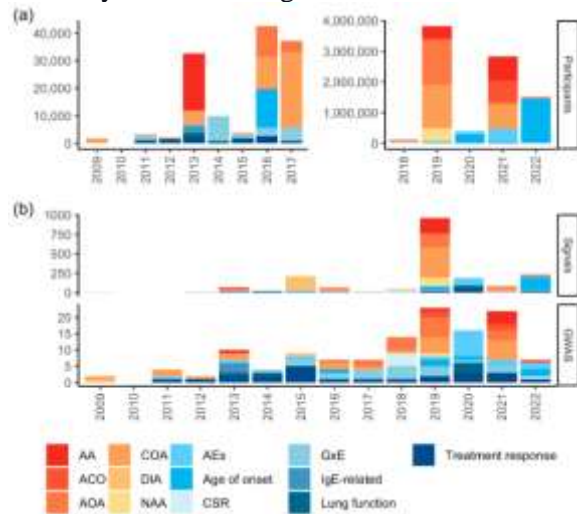


Fig 7 : Genomics of Treatable Traits in Asthma

7.2. Data Security and Privacy

The applications of big data for disease-related research will produce big privacy and security challenges to protect shared data and guarantee the privacy of individuals. Genomics studies conduct significance tests on large quantities of potentially sensitive data to understand the pattern of a disorder at the molecular level and to enable personalized treatments for a particular individual. In doing so, these studies have raised important security and privacy concerns and motivated serious consideration of the trade-offs between privacy risk and scientific and clinical benefits. This chapter surveys the proposed solutions to illustrate some of the trade-offs and to present their strengths and weaknesses. It is our opinion that the database infrastructure will need to provide secure computing capabilities in the search for associations between genetic markers and phenotypes. These capabilities should be applied selectively and audited for efficiency. In addition, we believe that regulations and collaborations should be able to handle the downstream processes of epidemiological investigations, including testing for the replication of association results and integration with other data sources. The current state of the art and open problems are summarized that have the promise of bridging the gap between the growing body of genetic and clinical data and powerful research programs. As a secondary contribution, this study simplifies and reconciles the earlier structuring of the challenges and solutions presented in the literature focusing on human genomics. This echoes similar efforts in the related areas of cryptographic constructions that would accommodate bioinformatics research and provide control over how the gene expression studies in which the biological samples are potentially identified.

Equation 3: Disease Prediction Based on Genetic Markers

Fig 8 : Disease–gene curation

The current scene in the analysis of disease is that its progress is in the area of managing, rather than resolving public health challenges. For example, the incidence of lifestyle diseases such as obesity, along with associated diseases of type 2 diabetes and some types of cancer, has been increasing rapidly throughout the world. Greater than 90% of those studied are a result of both genetic and environmental factors over the last 25 years. The genomics era's many studies have determined a set of genetic traits that increase susceptibility to these diseases with the help of a few environmental influencers. However, the association that exists between environmental and genetic factors rarely involves any direct interactions between the two. For such direct studies to occur, a very large, deep set of both environmental and genetic data would be needed to carry out associations across the selected research domains, and this is one of the major goals of future research. Complete and accurate generation of both sorts of data would allow meta-studies over not only different diseases across the globe but also the analysis of a much wider range of topics than that of current studies, leading to the development of new therapeutic approaches to tackle these complex diseases.

8. References

- [1] Syed, S. (2022). Breaking Barriers: Leveraging Natural Language Processing In Self-Service Bi For Non-Technical Users. Available at SSRN 5032632.
- [2] Nampally, R. C. R. (2022). Neural Networks for Enhancing Rail Safety and Security: Real-Time Monitoring and Incident Prediction. In *Journal of Artificial Intelligence and Big Data* (Vol. 2, Issue 1, pp. 49–63). Science Publications (SCIPUB). <https://doi.org/10.31586/jaibd.2022.1155>
- [3] Danda, R. R. (2022). Innovations in Agricultural Machinery: Assessing the Impact of Advanced Technologies on Farm Efficiency. In *Journal of Artificial Intelligence and Big Data* (Vol. 2, Issue 1, pp. 64–83). Science Publications (SCIPUB). <https://doi.org/10.31586/jaibd.2022.1156>
- [4] Rajesh Kumar Malviya , Shakir Syed , RamaChandra Rao Nampally , Valiki Dileep. (2022). Genetic Algorithm-Driven Optimization Of Neural Network Architectures For Task-Specific AI Applications. *Migration Letters*, 19(6), 1091–1102. Retrieved from <https://migrationletters.com/index.php/ml/article/view/11417>
- [5] Patra, G. K., Rajaram, S. K., Boddapati, V. N., Kuraku, C., & Gollangi, H. K. (2022). Advancing Digital Payment Systems: Combining AI, Big Data, and Biometric Authentication for Enhanced Security. *International Journal of Engineering and Computer Science*, 11(08), 25618–25631. <https://doi.org/10.18535/ijecs/v11i08.4698>
- [6] Syed, S. (2022). Integrating Predictive Analytics Into Manufacturing Finance: A Case Study On Cost Control And Zero-Carbon Goals In Automotive Production. *Migration Letters*, 19(6), 1078–1090.
- [7] Nampally, R. C. R. (2022). Machine Learning Applications in Fleet Electrification: Optimizing Vehicle Maintenance and Energy Consumption. In *Educational Administration: Theory and Practice*. Green Publication. <https://doi.org/10.53555/kuey.v28i4.8258>
- [8] Danda, R. R. (2022). Application of Neural Networks in Optimizing Health Outcomes in Medicare Advantage and Supplement Plans. *Journal of Artificial Intelligence and Big Data*, 2(1), 97–111. Retrieved from <https://www.scipublications.com/journal/index.php/jaibd/article/view/1178>
- [9] Chintale, P., Korada, L., Ranjan, P., & Malviya, R. K. (2019). Adopting Infrastructure as Code (IaC) for Efficient Financial Cloud Management. *ISSN: 2096-3246*, 51(04).
- [10] Kumar Rajaram, S.. AI-Driven Threat Detection: Leveraging Big Data For Advanced Cybersecurity Compliance. In *Educational Administration: Theory and Practice* (pp. 285–296). Green Publication. <https://doi.org/10.53555/kuey.v28i4.7529>
- [11] Syed, S. (2022). Leveraging Predictive Analytics for Zero-Carbon Emission Vehicles: Manufacturing Practices and Challenges. *Journal of Scientific and Engineering Research*, 9(10), 97–110.
- [12] RamaChandra Rao Nampally. (2022). Deep Learning-Based Predictive Models For Rail Signaling And Control Systems: Improving Operational Efficiency And Safety. *Migration Letters*, 19(6), 1065–1077. Retrieved from <https://migrationletters.com/index.php/ml/article/view/11335>

- [13] Danda, R. R. (2022). Deep Learning Approaches For Cost-Benefit Analysis Of Vision And Dental Coverage In Comprehensive Health Plans. *Migration Letters*, 19(6), 1103-1118.
- [14] Sarisa, M., Boddapati, V. N., Kumar Patra, G., Kuraku, C., & Konkimalla, S. (2022). Deep Learning Approaches To Image Classification: Exploring The Future Of Visual Data Analysis. In *Educational Administration: Theory and Practice*. Green Publication. <https://doi.org/10.53555/kuey.v28i4.7863>
- [15] Syed, S. (2022). Towards Autonomous Analytics: The Evolution of Self-Service BI Platforms with Machine Learning Integration. *Journal of Artificial Intelligence and Big Data*, 2(1), 84-96.
- [16] Nampally, R. C. R. (2021). Leveraging AI in Urban Traffic Management: Addressing Congestion and Traffic Flow with Intelligent Systems. In *Journal of Artificial Intelligence and Big Data* (Vol. 1, Issue 1, pp. 86–99). Science Publications (SCIPUB). <https://doi.org/10.31586/jaibd.2021.1151>
- [17] Ramanakar Reddy Danda. (2022). Telehealth In Medicare Plans: Leveraging AI For Improved Accessibility And Senior Care Quality.
- [18] Venkata Nagesh Boddapati, Manikanth Sarisa, Mohit Surender Reddy, Janardhana Rao Sunkara, Shravan Kumar Rajaram, Sanjay Ramdas Bauskar, Kiran Polimetla. Data migration in the cloud database: A review of vendor solutions and challenges . *Int J Comput Artif Intell* 2022;3(2):96-101. DOI: 10.33545/27076571.2022.v3.i2a.110
- [19] Syed, S. (2021). Financial Implications of Predictive Analytics in Vehicle Manufacturing: Insights for Budget Optimization and Resource Allocation. *Journal Of Artificial Intelligence And Big Data*, 1(1), 111-125.
- [20] Syed, S., & Nampally, R. C. R. (2021). Empowering Users: The Role Of AI In Enhancing Self-Service BI For Data-Driven Decision Making. In *Educational Administration: Theory and Practice*. Green Publication. <https://doi.org/10.53555/kuey.v27i4.8105>
- [21] Danda, R. R. (2021). Sustainability in Construction: Exploring the Development of Eco-Friendly Equipment. In *Journal of Artificial Intelligence and Big Data* (Vol. 1, Issue 1, pp. 100–110). Science Publications (SCIPUB). <https://doi.org/10.31586/jaibd.2021.1153>
- [22] Chandrakanth Rao Madhavaram, Eswar Prasad Galla, Hemanth Kumar Gollangi, Gagan Kumar Patra, Chandrababu Kuraku, Siddharth Konkimalla, Kiran Polimetla. An analysis of chest x-ray image classification and identification during COVID-19 based on deep learning models. *Int J Comput Artif Intell* 2022;3(2):86-95. DOI: 10.33545/27076571.2022.v3.i2a.109
- [23] Bauskar, S. R., Madhavaram, C. R., Galla, E. P., Sunkara, J. R., & Gollangi, H. K. (2022). PREDICTING DISEASE OUTBREAKS USING AI AND BIG DATA: A NEW FRONTIER IN HEALTHCARE ANALYTICS. In the *European Chemical Bulletin*. Green Publication. <https://doi.org/10.53555/ecb.v11:i12.17745>
- [24] Wang, X., & Thompson, B. . AI-Driven Insights: The Role of Cloud Analytics in Shaping Data Narratives for Business Strategy. *International Journal of Artificial Intelligence and Data Analytics*, 41(3), 184-203. <https://doi.org/10.1007/ijai..41.03>
- [25] Eswar Prasad Galla.et.al. (2021). Big Data And AI Innovations In Biometric Authentication For Secure Digital Transactions *Educational Administration: Theory and Practice*, 27(4), 1228 –1236 Doi: 10.53555/kuey.v27i4.7592
- [26] Roberts, J., & Garcia, P. (2022). Exploring AI Storytelling: Bridging the Gap Between Cloud Analytics and Actionable Insights. *Journal of Digital Transformation*, 8(1), 46-58. <https://doi.org/10.1016/j.jdt.2022.01.008>
- [27] Janardhana Rao Sunkara, Sanjay Ramdas Bauskar, Chandrakanth Rao Madhavaram, Eswar Prasad Galla, Hemanth Kumar Gollangi, Data-Driven Management: The Impact of Visualization Tools on Business Performance, *International Journal of Management (IJM)*, 12(3), 2021, pp. 1290-1298. <https://iaeme.com/Home/issue/IJM?Volume=12&Issue=3>
- [28] Lin, H., & Jackson, E.. AI-Driven Narrative Visualization: A New Era of Cloud-Based Data Storytelling. *Journal of Advanced Analytics*, 39(5), 415-431. <https://doi.org/10.1109/jaa..39.05>
- [29] Gagan Kumar Patra, Chandrababu Kuraku, Siddharth Konkimalla, Venkata Nagesh Boddapati, Manikanth Sarisa, An Analysis and Prediction of Health Insurance Costs Using Machine Learning-Based Regressor Techniques, *International Journal of Computer Engineering and Technology (IJCET)* 12(3), 2021, pp. 102-113. <https://iaeme.com/Home/issue/IJCET?Volume=12&Issue=3>
- [30] Chen, Z., & Patel, V. (2021). Storytelling with Data: Integrating AI into Cloud Analytics for Business Intelligence. *International Journal of Data Science*, 23(3), 98-115. <https://doi.org/10.1093/ijds.2021.23.03>

- [31] Venkata Nagesh Boddapati, Eswar Prasad Galla, Janardhana Rao Sunkara, Sanjay Ramdas Bauskar, Gagan Kumar Patra, Chandrababu Kuraku, Chandrakanth Rao Madhavaram, 2021. "Harnessing the Power of Big Data: The Evolution of AI and Machine Learning in Modern Times", *ESP Journal of Engineering & Technology Advancements*, 1(2): 134-146.
- [32] Miller, R., & Huang, W. . AI-Powered Narratives: Leveraging Cloud Analytics for Smarter Decision Making and Business Solutions. *Jour*
- [33] Ravi Kumar Vankayalapati , Chandrashekar Pandugula , Venkata Krishna Azith Teja Ganti , Ghatoth Mishra. (2022). AI-Powered Self-Healing Cloud Infrastructures: A Paradigm For Autonomous Fault Recovery. *Migration Letters*, 19(6), 1173–1187. Retrieved from <https://migrationletters.com/index.php/ml/article/view/11498>
- [34] Tulasi Naga Subhash Polineni , Kiran Kumar Maguluri , Zakera Yasmeen , Andrew Edward. (2022). AI-Driven Insights Into End-Of-Life Decision-Making: Ethical, Legal, And Clinical Perspectives On Leveraging Machine Learning To Improve Patient Autonomy And Palliative Care Outcomes. *Migration Letters*, 19(6), 1159–1172. Retrieved from <https://migrationletters.com/index.php/ml/article/view/11497>
- [35] Venkata Obula Reddy Puli, & Kiran Kumar Maguluri. (2022). Deep Learning Applications In Materials Management For Pharmaceutical Supply Chains. *Migration Letters*, 19(6), 1144–1158. Retrieved from <https://migrationletters.com/index.php/ml/article/view/11459>
- [36] Maguluri, K. K., Pandugula, C., Kalisetty, S., & Mallesham, G. (2022). Advancing Pain Medicine with AI and Neural Networks: Predictive Analytics and Personalized Treatment Plans for Chronic and Acute Pain Managements. In *Journal of Artificial Intelligence and Big Data* (Vol. 2, Issue 1, pp. 112–126). Science Publications (SCIPUB). <https://doi.org/10.31586/jaibd.2022.1201>
- [37] Lekkala, S. (2021). Ensuring Data Compliance: The role of AI and ML in securing Enterprise Networks. In *Educational Administration: Theory and Practice*. Green Publication. <https://doi.org/10.53555/kuey.v27i4.8102>