Original Article

Predictive Modeling of Carcinogen Exposure Using Big Data Analytics

Seon-woo Kim[1*]

Access this article online

Quick Response Code:

Website:
www.carcinogenesis.com

DOI:
10.4103/jcar.jcar_21_02_06

[1] Department of Pathology, College of Korean Medicine, Gachon University, Seong-nam, Republic of Korea

Address for correspondence:
Seon-woo Kim,
Department of Pathology,
College of Korean Medicine,
Gachon University, Seong-nam, Republic of Korea,

Submitted: 26-Sep-2022
Revised: 30-Oct-2022
Accepted: 8-Nov-2022
Published: 17-Dec-2022

Abstract

This study investigates the revolutionary potential of predictive modeling in identifying and limiting carcinogen exposure to improve cancer prevention techniques. Our study explores the complex interactions between environmental, occupational, and lifestyle variables to create predictive models that estimate the risks related to carcinogen exposure. We do this by using the broad field of big data analytics. The geographical dimension is one key factor that allows for the personalization of interventions based on regional differences in environmental features and industrial landscapes. Occupational environments, which are frequently high-risk ones, are examined to find patterns of exposure that may be used to influence specific occupational health and safety regulations. Evaluations of the air and water quality provide important insights that help legislators adopt exact pollution control measures and shape sustainable urban development. Lifestyle variables provide individualized risk evaluations since they are deeply integrated into predictive modeling. This multifaceted investigation provides guidance for population-wide interventions and individual preventative techniques, pushing public health policies toward proactive measures. The appropriate use of big data requires careful consideration of data privacy and ethical issues. Robust ethical frameworks maintain a difficult balance between preserving individual privacy and extracting significant insights, ensuring that scientific discoveries are realized with the highest regard for ethical principles. Predictive modeling is used in many fields, such as research and development, environmental management, public health policy, occupational health and safety, cancer prevention techniques, and healthcare resource allocation. According to this study, big data analytics and predictive modelling played crucial roles in the fight against cancer in the future, ushering in a new era of proactive, evidence-based treatments for a stronger, healthier society.

Keywords:

Predictive Modeling (PM), Carcinogen Exposure (CE), Big Data Analytics (BDA), Smart PLS Algorithm.

Introduction

A mathematical process that is used to foretell about the events of the future or any information about knowledge by the observation of historical data that is related to it. It is a statistical technique that is used for the prediction of future behavior.

It is a type of data mining technology in which historical and current data are compared, and a model is generated which gives ideas about future outcomes, which is known as predictive modeling. An analysis that is based on statistics performed by computers when input is given by operators[1]. The incorporation of state-of-the-art technology has created previously unheard-of present an enticing look into a future where we may actively reduce the dangers connected to carcinogenic compounds as we stand on the cusp of a new age in environmental science and healthcare. Being a powerful enemy in the field of public health, cancer has inspired scientists and researchers to investigate novel theories in order to comprehend its causes and create efficient preventative measures. Conventional approaches to researching carcinogen exposure frequently rely on retrospective studies that highlight past data and known risk factors. However, the advent of big data analytics has changed everything, enabling us to explore real-time, extensive datasets that capture a wide range of factors impacting exposure to carcinogens. Expected scenarios are generated from this data collection.

For reprints contact: editor@carcinogenesis.com

How to cite this article: Kim S-w. Predictive Modeling of Carcinogen Exposure Using Big Data Analytics. J Carcinog 2022; 21(2):44-51

Its applications are found in any industry, enterprise, or endeavor where data is collected. There are three types of predictive modeling which are discussed next. The first is a decision tree, which is very simple but extremely powerful in which many variables are analyzed. The second method is regression which is a compelling method in the field of statistics. And the third one is neural networks[2]. During the discussion of predictive carcinogenicity modeling, one of the significant steps is to observe that what the method of representation of carcinogens is so that better relationships and patterns can be seen in the data which will result in accuracy and generalization. Normally, two sources are considered authentic to know the carcinogenic nature of any substance. The first one is the laboratory test the second thing is epidemiology research that is related to research in people. In the most recent study, it is concluded that there are 63 carcinogenic substances[3].

Predictive modeling and big data analytics are married at the core of this transformation. With its foundations in statistical analysis and machine learning, predictive modeling allows us to identify patterns, correlations, and trends in enormous datasets. These models can provide a detailed knowledge of the dynamics of carcinogen exposure by using algorithms to decipher intricate interactions between environmental, occupational, and lifestyle variables. It is impossible to exaggerate the importance of predictive modeling when discussing exposure to carcinogens. In addition to helping us locate possible exposure sources, it is easier to create focused treatments and preventative measures. In public health, the capacity to predict and measure the risk of carcinogenic exposure at the individual and population levels is revolutionary, allowing researchers, policymakers, and medical professionals to take preventative measures that can dramatically lower the incidence of cancer. In the field of big data analytics, a foundation is set for the predictive modeling journey. Big data is a vast ocean of information drawn from several sources. It is distinguished by its volume, pace, and diversity. This enormous data bank is augmented by lifestyle surveys, health databases, occupational records, and environmental monitoring stations, providing a thorough foundation for developing predictive models. Combining data from many sources improves forecast accuracy and dependability while also adding value to the modeling process. Another report also shows that 163 substances are also responsible for causing cancer. Mutations may occur due to chemical agents which increase the chance of malignant transformation that become the cause of tumor formation. Such mutation becomes the cause of genetic damage before and then causes the progression in the development of cancer. There are also chances of increment in acquiring the characters of malignant tumors by benign tumors. A carcinogen is any substance, organism, or agent capable of inducing cancer[4]. There are two ways of carcinogen generation in the environment. The first one is naturally in the environment due to ultraviolet rays in the sunlight or many viruses are also responsible for them. The second source is artificial, which is created by humans by the use of automobile exhaust fumes and the use of cigarettes. Big data analytics involve the collection, examination, and analysis of large amounts of information to reveal trends of markets, insights, and patterns, which will prove helpful for the companies in the development of good decisions about business[5]. This important function is performed by a big data analyst who will analyze and uncover beneficial data like those trends and patterns that are not well known but hidden so that companies can find it new and make decisions about their business by considering that which will become the cause of competitive usefulness.The methodical understanding of cancer genomic data in the information about tumor biology and the meditative opportunities always remain demanding[6]. A robust preclinical model system provides a vast favor to represent the information about the genomic diversity of human cancer and a detailed description is available about genetic and pharmacology. The discussion topic is cancer cell line Encyclopedia, in which detailed information about gene expression and chromosomal copy number is found. The most important factors that are responsible for lung cancer are smoking, cigarette, and many other products of tobacco. If a person uses secondhand tobacco, then he will be in more danger zone than a person using fresh tobacco products[7]. But exposure to all these carcinogen-containing products, radiation, and many other indoor and outdoor pollution will become the beginning of tumor formation within the body which will lead to the death of the human body. In developing countries, the spread of smoking is going to increase day by day, which will become the cause of new lung cancer epidemics in these nations. A family history with a positive record and having symptoms of lung disease will represent the symptoms that are clinically risk indicators[8]. So it is concluded that all the deaths that happen due to lung cancer, are related to cigarette smoking.

This condition emphasizes the continuous efforts that are going on worldwide to control tobacco use. To have complete information about the elements responsible for lung cancer, one should understand the combined results that reveal the interrelationship between the two main causes. The first cause is exposure to etiologic agents and the second reason is susceptibility of every individual to these agents. A harmonious interaction between all the factors responsible for lung cancer may become the reason for considerable important consequences for lung cancer risk[9]. Data privacy and ethical issues become increasingly important as we learn more about predictive modeling for carcinogen exposure. In order to use big data responsibly, one must carefully strike a balance between gaining insightful knowledge and protecting people's privacy. Achieving this balance requires the creation of strong data governance guidelines and ethical frameworks that put protecting sensitive data first and promote scientific progress. A very popular example of this synergistic interaction is the synergistic effects that will be observed when a cigarette smoker is at risk of lung cancer due to continuous radiation exposure. No doubt, Sociopsychological, and environmental factors are responsible for lung cancer but certain factors that are innate in the host also increase the susceptibility of developing lung cancer. For example, a family that has a background of lung cancer will be at greater risk of developing lung cancer. All the causes that are responsible for developing cancer are strongly dependent on the potential of all cancer-causing agents whether they are strong or weak. To reduce the risk of lung cancer it is necessary to eliminate the exposure of population to the regular cancer-causing agents. So that there will be no chance of any tumor formation within the body that will cause cancer[10].

Research Objective

The main purpose of this study is to predict the effect of cancer-causing agents on a normal person and a person who already has a background of lung cancer. In this research, it was discussed how strongly one can be affected after exposure to cancer-causing agents.

Literature Review

Predictive Modeling of Carcinogen Exposure using Big Data Analytics

Carcinogens are those substances that may increase the risk of cancer exposure for human beings. Such substances might be physical as the form of ultraviolet rays directly from the sun such that in many countries Ozone layer has been depleted[11]. There are more than 100 of carcinogens that's increase the chances of cancers in humans. Various types of chemicals also cause the cancers such as the infections from various viruses. But, to have simple contact with the carcinogens does not develop cancer tissues in people[12].There is large amount of data that allows for the establishment of cancer on human bodies. Various work environments elevate the chances of cancer. Occupational carcinogens have a special place in various forms of carcinogens of humans. Up to 1970's the major amount of carcinogens are occupational until the recognition of non-occupational carcinogens[13].

Exposure to carcinogens have a harmful effect on the DNA structure of the humans. Some are DNA-damaging agents, it may be the exogenous agents but the endogenous processes have caused more damage. The breakage of the single and double strands break and covalent bond damage[14]. When predicting carcinogen exposure using predictive modeling, geographic factors are crucial. Variations in lifestyle patterns, industrial landscapes, and environmental features among different locations impact the kinds and incidence of carcinogens. By integrating spatial information into forecasting models, scientists may customize treatments according to the unique obstacles and hazards encountered by certain groups. This geographical component gives the models an extra degree of accuracy, guaranteeing that the recommendations generated are useful and appropriate for the given situation. The workplace is a major source of carcinogen exposure, and certain sectors are more likely to cause cancer than others because of the materials and procedures they utilize. Occupational data are considered via predictive modeling, which finds exposure patterns within certain occupations. This contributes to a safer work environment by informing occupational health and safety legislation and helping to identify and mitigate hazards for those employed in high-risk jobs. The presence of harmful chemicals in the fresh water fish may cause cancer as the fish are born with harmful chemicals that made them toxic to eat. Many key characteristics are present in substances causing the Cancer. Genotoxic and Immunosuppressive are some of the problems that originated from toxic cancer substances[15]. For the defense in the human body, the granulocytes chemically secrete the oxidants and radicals that fight back with the human genotoxic problems. This is an effective way to remove the harmful parasites from the human body. In the ordinary course of life, cells have enough guard systems to overcome the burden of inflammation in the human body but this happens at a negligible level. If this persistent inflammation increases the risk of cancer exposure [16]. In today era of industrial revolutions, industries release environmental endocrine disruptive chemicals (EDCs) such as pesticides and smoke that has disruptive effect on the human body.

The breathing system and lungs are affected. Most of the time, the effects of the chemicals have been observed on the animal's body, which disable the animals to develop the genetic tracts. It also affects the animals' and humans' behavior, metabolism and brain activities [17].Intake of Fresh juices and vegetables is positively associated with a decrease level of cancer in people in stomach, mouth, and cervix. Vitamin C is rarely used for the treatment of the cancer patients[18]. Many carcinogens affect the DNA Stem cells. It has negative effects on the development of the DNA models in the human body. The diseases that travel through DNA have harmful effect on the generations. If one generation has the cancer-causing features, the next generations may have chances of cancer[19].The dietary elements for a human must have those compounds that minimizes the chances of exposure to humans. Chemicals such as the phytochemicals lead to the minimization of the cancer problems[20]. Important elements affecting carcinogen exposure include the quality of the air and water, which are essential parts of our everyday surroundings. Data from satellite views, water quality evaluations, and environmental monitoring stations may all be integrated due to big data analytics. By identifying trends in pollution levels, predictive models can assist in identifying regions where there is a higher risk of exposure to carcinogens. Targeted initiatives to address environmental factors that contribute to cancer, such as pollution control policies and public awareness campaigns, are based on this knowledge. Lifestyle variables, which include eating patterns, exercise routines, and individual preferences, also have a major role in the exposure to carcinogens. In this perspective, surveys and health records that document personal lifestyles are essential parts of big data analytics.

By taking into account these lifestyle variables, predictive modeling illuminates the ways in which an individual's personal decisions interact with environmental and professional factors to determine their overall risk profile. With the digital era started growing there are multiple of environmental and population data which is compiled as giant. In today's worlds, new approaches of data mining, machine and deep learning helps in maintaining the large amount of data which helps in gaining the right decision in no time and with accurate evidences[5].The translation of the cancer genetic data into the information to be accessible at later time was the dilemma in past few years. With time the Cancer cell Line Encyclopedia (CCLE) in which the compilation of expressions of genes and copy numbers of the chromosomes have been defined and a huge amount of almost 947 human cancerous cells data is present. It helps pharmacological for the identification of the genes and their expressions[4]. Many a model have been developed for diagnosis of the breast cancer of the females. In today's environment, a large influx of the females has been suffering from the breast cancer problems keeping their lives at risk[3].

There might be the relationship between your DNA and carcinogens, DNA is present in the genes. Genes helps in making the proteins in your body which helps in the growth of the millions of cells in your body. When the carcinogens react with the genetic codes of the humans it changes the normal cells into the cancerous cells[21]. Sometimes the carcinogens directly hit the DNA of humans and halted its working and mutation is started. When the mutation starts, the DNA gives instruction to the cells to multiply with no limits which in turn transform itself into to the tumors or blood cancer. It is not the overnight process, but it takes many years to build the carcinogens in human body and develop the carcinogens cells[22]. The vast changes in the data modeling and data mining changing the techniques of the computations allows the clinicians to choose the best possible path from various available. The collection of data as "data mining" helps in the construction of the data in meaningful forms which helps in the detection of various predictable models in the cancerous humans[9]. The Lung cancer epidemic emerged in the mid-1900s. smoking of cigarettes is the main cause of the lungs cancer in humans which underrated all he efforts of the health communities from refraining the people from smoking. In developing nations, smoking is increased which hyped the chances of lungs cancer in people over there. With time risk prediction models have been developed which enables the clinicians and researchers to reduce the level of risk[8]. Collaboration between academics, decision-makers, medical professionals, and the general public is essential to realizing this objective. Predictive modeling has transformational potential, but it won't reach its full potential unless evidence-based solutions are put into practice, sustainable practices are promoted, and health is given priority in all areas of life. To sum up, the combination of big data analytics and predictive modeling signals a paradigm shift in how we interpret and reduce exposure to carcinogens. This multifaceted investigation gives us the ability to negotiate the complex network of lifestyle, work, and environmental factors, giving us a comprehensive understanding of the elements influencing cancer risk. Predictive modeling appears as a ray of hope as we approach the revelation of the future. It points us in the direction of a society where focused tactics and proactive interventions create a healthier, cancer-resistant environment. Ultimately, the predictive modelling of carcinogen exposure is a monument to our ability to use knowledge for the common good, particularly at this nexus of technology progress and public health. This is not the end of the journey; rather, it is a continuous process of discovery and application that shape a future where big data analytics and predictive modelling work together to prevent cancer before it becomes too late. In the contemporary business environment, the magnitude of data and its impact on the overall business environment has been increased. Business intelligence helps in maintaining the large amount of data[6]. Advance models of predictive carcinogens ae developed which reduces the risk of effects of harmful chemicals on human body and can be treated in time. Moreover, the DNA structures of humans can be protected with less damage[23].

Descriptive statistic

Table 1

Name	No.	Mean	Median	Scale min	Scale max	Standard deviation	Excess kurtosis	Skewness	Cramér-von Mises p value
PM 1	0	1.571	1.000	1.000	3.000	0.639	-0.477	0.692	0.000
PM2	1	1.490	1.000	1.000	3.000	0.610	-0.184	0.874	0.000
PM3	2	1.490	1.000	1.000	4.000	0.674	2.621	1.484	0.000
CE1	3	1.592	1.000	1.000	4.000	0.697	1.499	1.149	0.000
CE2	4	1.694	2.000	1.000	3.000	0.645	-0.664	0.403	0.000
CE3	5	1.408	1.000	1.000	2.000	0.491	-1.932	0.386	0.000

The above result present descriptive statistical analysis result describes that mean values, median values, the standard deviation rates, the skewness values of each variable.

the PM1 shows that mean value is 1.571 the standard deviation shows that 63% deviate from mean. The PM2, PM3 present that 49% average value the deviation rate is 61%, 67% respectively.

The overall probability value is 0.000 shows that 100% significantly level between them.

Table 2

	PM 1	PM2	PM3	CE1	CE2	CE3
CE1	-0.255	0.182	-0.009	1.000	0.000	0.000
CE2	0.078	-0.137	0.439	0.085	1.000	0.000
CE3	0.102	-0.054	-0.296	-0.109	-0.443	1.000
PM 1	1.000	0.000	0.000	0.000	0.000	0.000
PM2	-0.142	1.000	0.000	0.000	0.000	0.000
PM3	-0.034	-0.037	1.000	0.000	0.000	0.000

The above result describe that correlation coefficient analysis result present some negative and some positive link between them. the CE1 shows that -0.255 the PM2 shows that 18% significant analysis between them. the overall result present that significant link between them.

Significant analysis:

Table 3

Matrix	Original sample (O)	Sample Mean (M)	2.5%	97.5%
CE<-CE	0.003	0.056	-0.582	0.717
CE2<-CE	1.000	0.885	0.535	1.008
CE3<-CE3	1.000	1.000	1.000	1.000
PM1<-PM	0.164	0.089	-0.596	0.696
PM2<-PM	-0.291	-0.081	-0.646	0.693
PM3<-PM	0.930	0.587	-0.843	1.001

The above result describe that significant analysis result describe the original sample values, sample mean values, 2.5% rate also that 97.5% of each matrix. The first factors are CE<-CE its original sample value is 0.003 the mean value is 5% the 2.5% confidence interval rate is -0.582 the 97.5% confidence interval rate is 71% respectively.

The third matrix is CE3<-CE3 result shows that original sample value is 1.000 respectively of each indicators. The last matrix presents that PM3<-PM its rate of original sample value is 93% the mean value is 58% the 2.5% confidence interval rate is -0.843 and 97.5% confidence interval rate is 1.001 respectively.

Model Fitness:

Table 4

Factors	Saturated Model	Estimated Model
SRMR	0.117	0.131
d-ULS	0.288	0.360
d-G	0.058	0.074
Chi square	15.564	19.153
NFI	0.498	0.382

The above result describes that saturated model and estimated model the result present SRMR, d-ULS, D-G rates, chi square and NFI rates. The saturated model present that 0.117, 0.288, 0.058 the chi square value is 15.564 shows that positive chi square value of saturated model. The result present that NFI rate is 0.498 shows that 49% saturated model between them. the estimated model present that 13% SRMR value, the d-ULS rate is 36% the chi square value is 19.153 shows that positive chi square rate between them. the NFI rate of estimated model is 38% respectively.

Table 5

Variables	BIC (Bayesian information Criterion)
CE	-4.899
CE3	-3.915

The above result present that model summary also shows for analysis result present BIC rates of CE and CE3 are -4.899 and -3.915 respectively.

Smart PLS Algorithm Model:

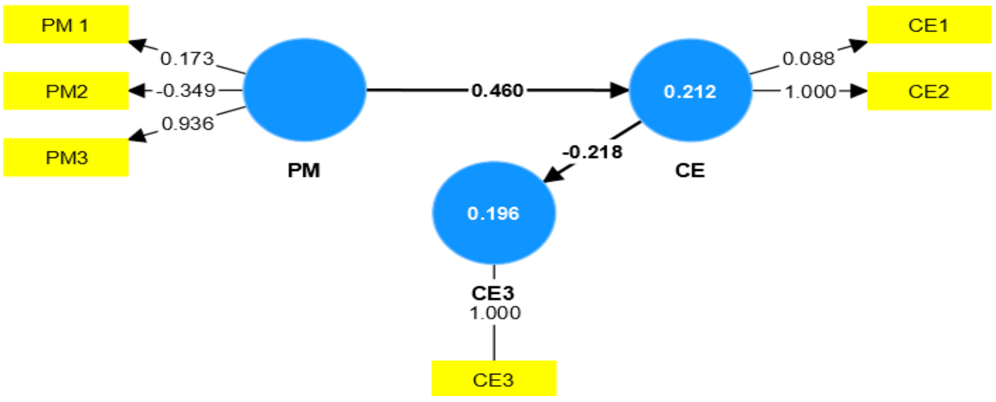

Figure 1

The above graph present that effects the PM shows that 0.173, 0.349, 0.936 positive points between the CE its rates are 0.173, 0.349, 0.936 all of them are present that positive rates. The PM shows that 46% significant relation between the CE and PM. The CE3 shows that negative link with CE its rate is -0.218 respectively.

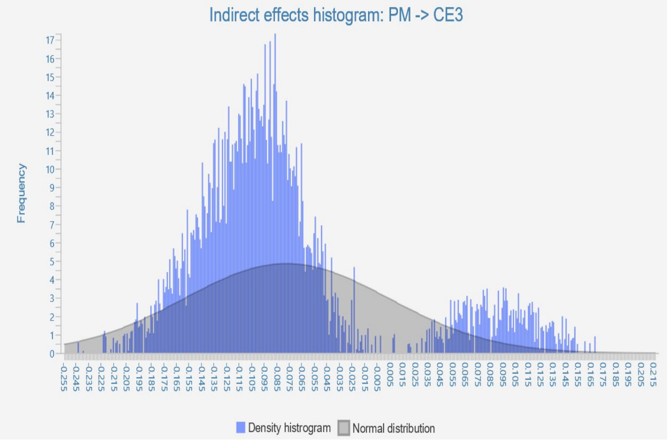

Figure 2

The above graph presents those indirect effects histogram the result present that frequency level its present at start point is 0 and end at 17 points. The horizontal side describe those negative points and some positive points the starting point is -0.255 and the end point is 0.215 respectively. The blue bar line shows that indirect effects histogram between them.

Applications:

Applications of big data analytics for predictive modeling of carcinogen exposure are numerous and significant, reaching across public health, environmental science, and policymaking areas. The following significant uses illustrate the method's transformational potential:

Strategies for Preventing Cancer:

• Individual Risk Assessment: By using predictive models, one may determine a person's personal risk of carcinogen exposure by taking into account their lifestyle, career, and place of residence. Specific tests or lifestyle changes are among the targeted preventative measures that are informed by this individualized risk assessment.

• Population-wide Interventions: Public health authorities can target groups or areas with greater expected risks with population-wide interventions by analyzing massive databases. This might entail putting environmental laws into effect, running public awareness campaigns, or improving access to medical care.

Safety and Health at Work:

• Risk Mitigation in High-Risk Professions: High-risk professions and particular exposure patterns within them can be identified with the use of predictive modelling. With the use of this data, customized occupational health and safety policies may be developed, guaranteeing that employees in high-risk settings have enough protection.

• Regulatory Compliance: Predictive models may be used by industries to evaluate and improve adherence to workplace safety and environmental requirements. This proactive strategy lowers the long-term health risks linked to occupational carcinogen exposure and makes the workplace a safer place to work.

Management of the Environment:

• Targeted Pollution Control Measures: Areas with high concentrations of carcinogenic pollutants can be identified with the use of predictive models that incorporate environmental data. Policymakers may use this information to help them adopt targeted pollution control methods, such better waste management practices and strategies for reducing emissions.

• Urban Planning and Zoning: Predictive modeling is a useful tool for city planners to guide zoning and urban development policies. This helps to create healthier living environments by ensuring that public places, industrial zones, and residential areas are intended to minimize carcinogenic exposures.

Policies for Public Health:

• Evidence-Based Policy Formulation: A strong evidence foundation for public health policy formulation is provided by predictive modelling. Policymakers may create preventative programs, lobby for laws targeted at lowering carcinogen exposure on a larger scale, and prioritize budget allocation using the information gained from these models.

• Monitoring and Surveillance: Real-time surveillance of trends in carcinogen exposure is made possible by the ongoing monitoring of prediction model outputs. This dynamic approach makes it possible to quickly modify policies and take action in response to new hazards or evolving trends in occupational and environmental exposures.

Allocation of Healthcare Resources:

• Resource Planning: Based on anticipated exposure levels, healthcare practitioners may utilize predictive models to foresee and prepare for possible increases in cancer cases. This makes it easier to allocate resources optimally and guarantees that healthcare institutions are ready to handle the changing requirements of the populace.

• Early Detection and Intervention: In high-risk locations, early detection programs may be implemented with the use of predictive models. By taking a proactive stance, the likelihood of an early diagnosis and intervention is increased, which ultimately improves treatment results and lessens the financial strain that cancer places on healthcare systems.

Investigation and Creation:

• developing Carcinogen Identification: By seeing patterns and trends in data, predictive modeling helps identify developing carcinogens. Researchers trying to comprehend the changing panorama of cancer risks and create plans to counter new dangers find this knowledge to be quite helpful.

• Drug Discovery: Predictive modeling may be used in the pharmaceutical industry to speed up drug discovery procedures. Researchers can create more potent therapeutic approaches for the treatment and prevention of cancer by knowing the precise carcinogenic pathways and targets. Essentially, big data analytics and predictive modeling of carcinogen exposure have implications in every aspect of public health and cancer prevention. With the application of sophisticated analytics to these fields, there is hope for a day when society's efforts will be proactive rather than reactive, reducing hazards before they materialize and laying the groundwork for a stronger, healthier community.

Conclusion

To sum up, the integration of big data analytics and predictive modeling is a novel approach in our joint endeavor to understand the intricacies of exposure to carcinogens. The ramifications for environmental management, cancer prevention, and public health are significant as we navigate this innovative terrain. With the use of predictive modeling and the large amounts of big data it contains, we can now grasp carcinogenic risk factors beyond conventional limits. Finding patterns and correlations in massive datasets offers a more nuanced view of the interactions between lifestyle, work, and environmental variables, forming a more thorough knowledge of the dynamics of carcinogen exposure. These models get additional accuracy from the geographical component, which allows for customized solutions that take into account the particular difficulties that various communities experience. Through the recognition and resolution of geographical disparities in environmental attributes and industrial configurations, predictive modeling transforms into an adaptable instrument for devising focused approaches to reduce the likelihood of cancer.

Predictive modeling applies to occupational environments, which are frequently hotspots for carcinogen exposure. These models support the development of evidence-based occupational health and safety standards in addition to protecting the health of those employed in high-risk occupations through the analysis of occupational records. The effect spreads outside the office, encouraging a preventative and awareness-based culture. The quality of air and water, which are essential components of our everyday environment, are highlighted in predictive modeling initiatives. These models may pinpoint locations with higher risks of carcinogen exposure by combining information from satellite views, environmental monitoring stations, and water quality evaluations. This information forms the basis for focused initiatives, influencing legislation to address environmental causes of cancer and promote environmentally friendly behaviors. Lifestyle variables provide an insight into the individual decisions determining carcinogen exposure and are included into predictive modelling through surveys and health records. Comprehending the interplay between personal decisions and occupational and environmental factors offers a comprehensive understanding of an individual's whole risk profile. This knowledge not only helps to tailor preventative care but also directs public health campaigns to encourage healthy living. data privacy and ethical issues are major concerns among the promises of big data analytics and predictive modelling. It is crucial to strike a careful balance between obtaining insightful information and protecting personal privacy. Building strong ethical frameworks and data governance laws is essential to the responsible use of big data, since it guarantees that scientific breakthroughs are made with the highest regard for ethical norms and privacy. A picture of a future in which targeted tactics and proactive treatments are the standard emerges when we consider the path via the predictive modelling of carcinogen exposure using big data analytics. This multifaceted investigation turns into a ray of hope, pointing us in the direction of a future in which the prevalence of cancer is reduced and the general well-being of our community is strengthened.

References

1. R. Sadiq and M. J. Rodriguez, "Disinfection by-products (DBPs) in drinking water and predictive models for their occurrence: a review," *Science of the total Environment,* vol. 321, no. 1-3, pp. 21-46, 2004.
2. "The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models," *Nature biotechnology,* vol. 28, no. 8, pp. 827-838, 2010.
3. J. Tyrer, S. W. Duffy, and J. Cuzick, "A breast cancer prediction model incorporating familial and personal risk factors," *Statistics in medicine,* vol. 23, no. 7, pp. 1111-1130, 2004.
4. J. Barretina *et al.*, "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature,* vol. 483, no. 7391, pp. 603-607, 2012.
5. C.-L. Chan and C.-C. Chang, "Big Data, Decision Models, and Public Health," vol. 19, ed: MDPI, 2022, p. 8543.
6. H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," *MIS quarterly,* pp. 1165-1188, 2012.
7. D. B. Rubin, "Estimating causal effects from large data sets using propensity scores," *Annals of internal medicine,* vol. 127, no. 8_Part_2, pp. 757-763, 1997.
8. A. J. Alberg, M. V. Brock, J. G. Ford, J. M. Samet, and S. D. Spivack, "Epidemiology of lung cancer: Diagnosis and management of lung cancer: American College of Chest Physicians evidence-based clinical practice guidelines," *Chest,* vol. 143, no. 5, pp. e1S-e29S, 2013.
9. R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: current issues and guidelines," *International journal of*

medical informatics, vol. 77, no. 2, pp. 81-97, 2008.

10. K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal,* vol. 13, pp. 8-17, 2015.
11. V. Bouvard *et al.*, "A review of human carcinogens—Part B: biological agents," *The lancet oncology,* vol. 10, no. 4, pp. 321-322, 2009.
12. J. Siemiatycki *et al.*, "Listing occupational carcinogens," *Environmental health perspectives,* vol. 112, no. 15, pp. 1447-1459, 2004.
13. T. Sugimura, "Nutrition and dietary carcinogens," *Carcinogenesis,* vol. 21, no. 3, pp. 387-395, 2000.
14. J. L. Barnes, M. Zubair, K. John, M. C. Poirier, and F. L. Martin, "Carcinogens and DNA damage," *Biochemical Society Transactions,* vol. 46, no. 5, pp. 1213-1224, 2018.
15. R. Baan *et al.*, "A review of human carcinogens—part F: chemical agents and related occupations," *The lancet oncology,* vol. 10, no. 12, pp. 1143-1144, 2009.
16. K. Z. Guyton *et al.*, "Application of the key characteristics of carcinogens in cancer hazard identification," *Carcinogenesis,* vol. 39, no. 4, pp. 614-622, 2018.
17. F. Fitzpatrick, "Inflammation, carcinogenesis and cancer," *International immunopharmacology,* vol. 1, no. 9-10, pp. 1651-1667, 2001.
18. S. S. Mirvish, "Effects of vitamins C and E on N-nitroso compound formation, carcinogenesis, and cancer," *Cancer,* vol. 58, no. S8, pp. 1842-1850, 1986.
19. S. Ohnishi *et al.*, "DNA damage in inflammation-related carcinogenesis and cancer stem cells," *Oxidative medicine and cellular longevity,* vol. 2013, 2013.
20. N. Khambete and R. Kumar, "Carcinogens and cancer preventors in diet," *International Journal of Nutrition, Pharmacology, Neurological Diseases,* vol. 4, no. 1, pp. 4-10, 2014.
21. P. K. Panda, S. Mukhopadhyay, D. N. Das, N. Sinha, P. P. Naik, and S. K. Bhutia, "Mechanism of autophagic regulation in carcinogenesis and cancer therapeutics," in *Seminars in cell & developmental biology*, 2015, vol. 39: Elsevier, pp. 43-55.
22. C. Franken *et al.*, "Environmental exposure to human carcinogens in teenagers and the association with DNA damage," *Environmental research,* vol. 152, pp. 165-174, 2017.
23. K. Ounanian *et al.*, "Conceptualizing coastal and maritime cultural heritage through communities of meaning and participation," *Ocean & Coastal Management,* vol. 212, p. 105806, 2021.