Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Harnessing Big Data and Predictive Analytics for Early Detection and Cost Optimization in Cancer Care

Md Habibur Rahman[1], Md Ekrim Hossin[2], Md Jubayar Hossain[3], Syed Mohammed Muhive Uddin[4], Md Imtiaz Faruk[5], Md Mazharul Anwar[6] and Forhad Hossain[7] ✉

[12]School of Business, International American University, Los Angeles, CA 90010, USA
[35]Department of Management & Information Technology, St. Francis College, Brooklyn, NY, USA
[4]Department of Business Administration, Washington University of Science and Technology, Alexandria, VA 22314, USA
[67]Department of Statistics and Data Science, Jahangirnagar University, Savar, Bangladesh

Corresponding Author: Forhad Hossain, **E-mail**: forhadhossain.ju97@gmail.com

| ABSTRACT

Cancer continues to be a significant worldwide health concern, with increasing incidence and death imposing substantial expenses on patients, healthcare providers, and country economies (Bray et al., 2021; Sung et al., 2021). Conventional diagnostic and therapeutic approaches frequently do not identify cancers at an early, more manageable stage, while rising healthcare expenditures provide sustainability issues for both affluent and resource-limited healthcare systems (Mariotto et al., 2020; Yabroff et al., 2021). The emergence of big data and predictive analytics presents significant potential in tackling these dual concerns by facilitating earlier cancer detection and optimizing costs throughout the continuum of care. Big data in cancer includes diverse information sources such as electronic health records (EHRs), genomic databases, image archives, wearable devices, and insurance claims (Chen & Zhang, 2014; Raghupathi & Raghupathi, 2014). Predictive analytics, utilizing machine learning (ML), artificial intelligence (AI), and statistical modeling, enables doctors to discern concealed trends, anticipate illness progressions, and tailor treatment protocols (Obermeyer & Emanuel, 2016; Esteva et al., 2019). This study examines how the utilization of big data and predictive analytics might transform cancer care, emphasizing their functions in early detection and cost efficiency. The study conducts a thorough literature review and employs a methodological framework to analyze case examples from breast, lung, colorectal, and cervical cancers, illustrating the effectiveness of predictive models in early malignancy detection, minimizing late diagnoses, and facilitating cost-efficient interventions (Kourou et al., 2015; Cruz & Wishart, 2006; Topol, 2019). Moreover, it underscores the capacity of predictive modeling to reduce superfluous processes, optimize resource allocation, and enhance the value-based care framework (Yu et al., 2018; Rathore et al., 2017). Significant issues, such as data privacy, algorithmic bias, regulatory obstacles, and disparities in access, are examined to highlight the ethical and societal ramifications of this developing paradigm. The research contends that the incorporation of predictive analytics into oncology workflows is both a scientific need and an economic necessity, paving the way for precision oncology that improves survival rates while maintaining financial viability.

| KEYWORDS

Harnessing Big Data; Predictive Analytics; Early Detection; Cost Optimization; Cancer Care

| ARTICLE INFORMATION

ACCEPTED: 30 December 2024 **PUBLISHED:** 31 December 2024 **DOI:** 10.32996/jcsts.2024.6.5.22

1. Introduction

Cancer has become a significant health concern of the 21st century, with approximately 19.3 million new cases and 10 million fatalities worldwide in 2020 (Sung et al., 2021). With aging populations, evolving lifestyles, and increasing environmental exposures, the worldwide cancer incidence is anticipated to reach around 30 million new cases per year by 2040 (Bray et al.,

2021). Cancer not only has clinical implications but also exerts significant economic pressure; in the United States, cancer care expenditures surpassed $200 billion in 2020 and are anticipated to increase due to escalating incidence, costly treatments, and extended survivorship (Mariotto et al., 2020; Yabroff et al., 2021). Low- and middle-income nations encounter supplementary challenges stemming from insufficient screening infrastructure, restricted diagnostic capabilities, and disjointed healthcare systems (Ginsburg et al., 2017). Consequently, the simultaneous problems of enhancing early detection and reducing expenses constitute a global necessity for cancer management.

Early identification is widely acknowledged as fundamental to efficient cancer care, since malignancies identified at initial stages typically exhibit improved prognoses, broader therapy alternatives, and reduced management expenses (Etzioni et al., 2003; Richards, 2009). Conventional screening methods—such as mammography for breast cancer, Pap smears for cervical cancer, and colonoscopy for colorectal cancer—are constrained by inconsistent participation, false positives and negatives, and logistical difficulties (Duffy et al., 2017; Smith et al., 2019). Furthermore, nascent tumors including pancreatic and ovarian malignancies frequently lack dependable early detection methods, resulting in late-stage diagnoses and decreased survival rates (Siegel et al., 2022). In this perspective, big data and predictive analytics signify a transformative transformation. Utilizing extensive amounts of diverse data—including genetic and proteomic markers, imaging radiomics, and longitudinal health records—predictive models can detect nuanced indicators of malignancy well in advance of conventional diagnostics (Kourou et al., 2015; Topol, 2019).

The term "big data" in healthcare denotes datasets distinguished by substantial volume, rapid velocity, diverse variety, and high veracity (Laney, 2001; Raghupathi & Raghupathi, 2014). In oncology, this includes organized clinical data (diagnoses, medications, laboratory results), unstructured clinical notes, high-dimensional omics datasets (DNA, RNA, proteomics, metabolomics), medical imaging, and patient-generated health data from wearable devices (Chen & Zhang, 2014; Jensen et al., 2012). Predictive analytics use computational models, such as regression analysis, Bayesian inference, machine learning, and deep learning algorithms, to derive insights and anticipate clinical outcomes (Obermeyer & Emanuel, 2016; Esteva et al., 2019). These instruments facilitate the change from reactive to proactive oncology care, emphasizing the prediction and prevention of disease progression rather than only treating advanced-stage tumors (Collins & Varmus, 2015).

Besides clinical advantages, predictive analytics presents considerable potential for cost optimization. Expenditures on cancer care are influenced not just by innovative immunotherapies and precision medicine but also by inefficiencies, including superfluous diagnostic tests, preventable hospitalizations, and unjustified treatment variability (Yu et al., 2018; Sorenson et al., 2013). Predictive models can categorize patients by risk, enhance care paths, and avert resource overutilization. Hospital readmission prediction algorithms have been utilized in oncology to minimize avoidable admissions and related expenses (Rathore et al., 2017). Predictive cost models can assist insurers and governments in designing value-based payment models that link incentives with patient outcomes (Porter, 2010; Institute of Medicine, 2013).

Notwithstanding their potential, the application of big data and predictive analytics in oncology poses significant hurdles. Data silos, interoperability deficiencies, and uneven standards hinder the integration of varied datasets (Murdoch & Detsky, 2013; Belle et al., 2015). Ethical and regulatory issues—such as patient privacy, consent, algorithmic transparency, and bias—must be meticulously addressed to foster confidence and guarantee equitable implementation (Price & Cohen, 2019; Mittelstadt et al., 2016). Moreover, gaps in access to sophisticated analytics technology threaten to intensify existing inequities in cancer care across high- and low-resource environments (Kohli & Tan, 2016).

This study intends to rigorously analyze how leveraging big data, and predictive analytics might fulfill the dual objectives of early detection and cost minimization in cancer treatment. The literature review consolidates information from current uses of big data and predictive analytics in oncology, while the methodological framework delineates strategies for integrating various data sources and analytical tools. Applications are examined via case studies spanning various cancer types, subsequently evaluating the function of predictive modeling in cost optimization along the cancer treatment continuum. Ethical dilemmas and prospective research avenues are examined to offer a comprehensive perspective on the facilitators and impediments to adoption. This study emphasizes the transformative potential of big data and predictive analytics within clinical and economic frameworks to enhance survival outcomes and assure the financial sustainability of healthcare systems worldwide.

2. Literature Review

The combination of big data and predictive analytics in oncology has emerged as a critical area in medical research, mirroring larger trends in digital health transformation and precision medicine. Big data, defined by volume, velocity, diversity, and veracity (Laney, 2001), is ideally suited to the complexities of cancer care, where genetic, imaging, clinical, and behavioral information intersect. Predictive analytics, which includes statistical modeling, machine learning (ML), and artificial intelligence (AI), converts these diverse data streams into actionable insights for early detection, treatment personalization, and cost management (Obermeyer & Emanuel, 2016; Kourou et al., 2015). This review synthesizes the literature across three domains: (i) the importance

of big data in healthcare and oncology, (ii) predictive analytics for early cancer detection, and (iii) evidence on predictive modeling for cost optimization in cancer care.

2.1 Big Data in Healthcare and Oncology

Healthcare produces some of the world's most data-rich environments, thanks to electronic health records (EHRs), imaging modalities, genetic sequencing, insurance claims, and patient-generated health data (Raghupathi and Raghupathi, 2014; Belle et al., 2015). Oncology is particularly data-intensive because it combines multi-omics datasets (DNA, RNA, proteomics, metabolomics), radiological and histological imaging, clinical trial data, and longitudinal follow-up records (Chen & Zhang, 2014; Jensen et al., 2012). The Human Genome Project and next-generation sequencing technology have fueled the sequencing revolution, significantly expanding genomic databases important to cancer biology and allowing for the identification of mutations, biomarkers, and therapeutic targets (Collins & Varmus, 2015; Garraway, 2013). Similarly, advancements in digital pathology and radiomics have resulted in high-dimensional imaging datasets that can capture tumour heterogeneity (Gillies et al., 2016).

Big data's potential in oncology is based on its capacity to integrate many information sources into comprehensive patient profiles. Structured data, such as diagnoses, medications, and laboratory results, are captured by EHRs, whereas unstructured data, such as clinical notes, pathology reports, and social determinants of health, contain latent insights that can be extracted using natural language processing (Murdoch & Detsky, 2013; Jensen et al., 2012). Patient-generated health data from wearable devices and mobile apps enable continuous monitoring of activity, vital signs, and medication adherence, providing real-time insights into patient well-being (Wang et al., 2018). Insurance claims and registry data broaden the area of study to include healthcare consumption, cost trends, and population-level patterns (Miksad and Abernethy, 2018). Oncology can shift from reactive interventions to proactive, data-driven care models by collecting and evaluating data from these sources.

2.2 Predictive Analytics for Early Detection of Cancer

The early identification of cancer is a crucial factor influencing survival outcomes. Conventional screening techniques, including mammography, Pap smears, colonoscopy, and low-dose CT, have lowered mortality rates for breast, cervical, colorectal, and lung cancers, respectively (Smith et al., 2019; Duffy et al., 2017). However, these modalities encounter ongoing challenges: minimal adoption in marginalized populations, inaccuracies resulting in false positives and negatives, and constraints in identifying physiologically aggressive or uncommon malignancies (Etzioni et al., 2003; Richards, 2009). Predictive analytics improves early detection by utilizing extensive datasets and computational algorithms to identify high-risk individuals, recognize subtle disease indicators, and refine screening procedures.

Machine learning and artificial intelligence methodologies have shown considerable potential in oncology diagnoses. Convolutional neural networks (CNNs) and deep learning algorithms have been utilized in mammography, CT, and MRI scans, surpassing radiologists in certain contexts by decreasing false positives and enhancing sensitivity (McKinney et al., 2020; Esteva et al., 2019). Google Health's AI-driven breast cancer screening system demonstrated enhanced precision in detecting cancers among varied populations (McKinney et al., 2020). Radiomics, which involves the extraction of quantitative information from medical pictures, in conjunction with machine learning, has facilitated non-invasive identification of tumor subtypes, aggressiveness, and treatment response (Aerts et al., 2014; Gillies et al., 2016).

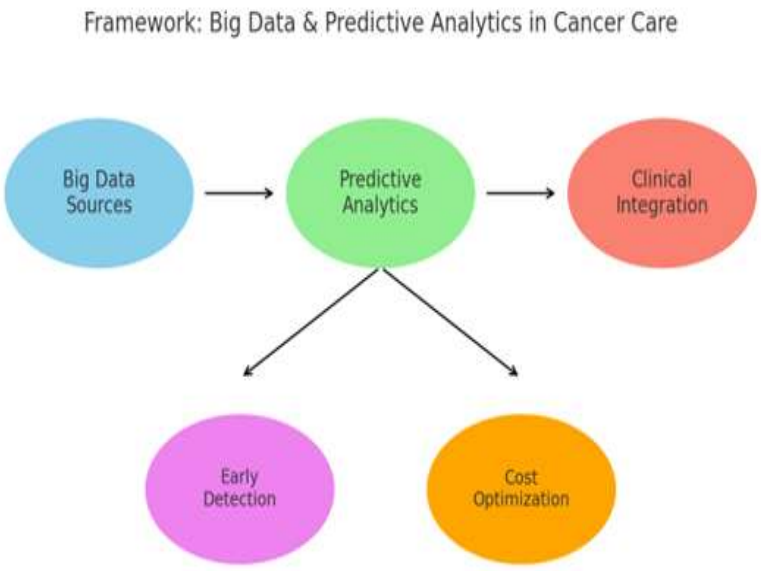

Figure 1. Conceptual framework of big data and predictive analytics in cancer care

In genomics, predictive analytics has proved crucial for detecting hereditary cancer risk. Polygenic risk scores (PRS) obtained from genome-wide association studies (GWAS) facilitate the stratification of populations for malignancies like breast, prostate, and colorectal (Mavaddat et al., 2019; Schumacher et al., 2018). Machine learning algorithms enhance predicted accuracy by incorporating polygenic risk scores with lifestyle, demographic, and clinical data (Inouye et al., 2018). Novel liquid biopsy technologies that identify circulating tumor DNA (ctDNA) and exosomes utilize predictive algorithms to differentiate early-stage cancer indicators from background interference, providing minimally invasive detection methods (Cohen et al., 2018; Wan et al., 2017).

In addition to individual applications, predictive analytics facilitates population-level cancer surveillance. Predictive models utilizing demographic, environmental, and behavioral data have been employed to project cancer incidence trends, assisting public health organizations in focusing screening initiatives (Mehta et al., 2019; Yu et al., 2018). Furthermore, predictive analytics improves individualized screening by customizing frequency and modality according to individual risk profiles instead of implementing standardized standards, thus enhancing efficiency and minimizing unnecessary procedures (Esserman et al., 2009; Topol, 2019).

2.3 Predictive Modeling for Cost Optimization

Cancer has a significant financial impact, with expenses covering palliative care, treatment, screening, prevention, diagnosis, and survivorship care. Drugs and hospital stays were the main causes of the more than $200 billion in healthcare costs associated with cancer in the United States in 2020 (Mariotto et al., 2020; Yabroff et al., 2021). Clinical outcomes and resource allocation must be balanced for oncology cost optimization, and predictive analytics has being used more and more to accomplish this.

In order to help payers and providers better spend resources, predictive models have been developed to estimate treatment costs, hospital readmissions, and healthcare use (Rathore et al., 2017; Sorenson et al., 2013). For instance, high-risk groups have been identified using risk prediction models for hospital readmissions among cancer patients, allowing for focused treatments that lower preventable admissions and related expenses (Ong et al., 2017). Predictive modeling can also be used to evaluate the

cost-effectiveness of new interventions, which can help policymakers make decisions about coverage and payment (Grosse et al., 2008; Neumann et al., 2018).

Predictive analytics is especially well-suited to value-based care systems, which prioritize results over expenditures. Predictive models assist accountable care organizations (ACOs), bundled payment plans, and oncology care models that align incentives for efficiency and quality by classifying patients by clinical and financial risk (Porter, 2010; Yu et al., 2018). Moreover, clinical decision support systems (CDSS) have incorporated predictive cost models to help oncologists choose treatment plans that strike a compromise between price and effectiveness (Spratt et al., 2018).

Predictive analytics has larger economic implications in addition to direct healthcare expenditures. Indirect costs of cancer care may include long-term impairment, caregiver stress, and lost productivity (Bradley et al., 2011). The economic impact of cancer survivorship has been estimated using predictive workforce analytics, which has also been used to guide workplace practices that reduce productivity losses (Foysal Mahmud et al., 2024). By include these indirect costs in predictive models, thorough evaluations of the financial burden of cancer are improved, and more sustainable approaches to healthcare funding are informed (Ginsburg et al., 2017).

2.4 Gaps and Limitations

Despite expanding data, there are still several gaps in the research. First, many predictive analytics models in oncology are based on single-institution or homogeneous datasets, restricting their applicability to varied populations (Rajkomar et al., 2019; Beam & Kohane, 2018). Algorithmic bias is still an issue, as underrepresentation of minority groups might worsen health disparities in early detection and care optimization (Obermeyer et al., 2019; Chen et al., 2020). Second, interoperability and data standardization issues impede integration of multimodal datasets from EHRs, genomics, and imaging (Murdoch & Detsky, 2013; Belle et al., 2015). Third, while many studies show predictive accuracy in controlled research settings, translation into real-world clinical workflows is still limited due to regulatory, reimbursement, and infrastructure hurdles (Price & Cohen, 2019; Kohli & Tan, 2016).

Finally, cost-effectiveness studies using predictive analytics in oncology are still in their early stages, with little robust evidence of return on investment (ROI). Few large-scale randomized controlled trials (RCTs) have thoroughly tested whether predictive models save costs while maintaining patient outcomes (Neumann et al., 2018; Spratt et al., 2018). Addressing these gaps necessitates multicenter collaborations, federated learning methodologies that protect privacy while pooling data, and thorough economic assessments that link predictive analytics to long-term cancer therapy.

3. **Methodological Framework**

The methodological foundation for using big data and predictive analytics in cancer care necessitates a multifaceted strategy that includes data collection, preprocessing, analytical modeling, validation, interpretation, and clinical integration. Each level presents unique technical, ethical, and organizational problems, but it also offers opportunity for dramatic advances in early detection and cost efficiency.

3.1 Data Sources in Oncology

The collection of data from many sources is a critical component of predictive analytics in cancer care. Electronic health records (EHRs) are the foundation of cancer big data, providing longitudinal patient information such as demographics, diagnoses, procedures, drugs, and laboratory findings (Murdoch & Detsky, 2013; Raghupathi & Raghupathi, 2014). However, EHRs frequently contain unstructured clinical notes, radiological reports, and pathology narratives that require natural language processing (NLP) to extract meaningful insights (Jensen et al., 2012; Wang et al., 2018).

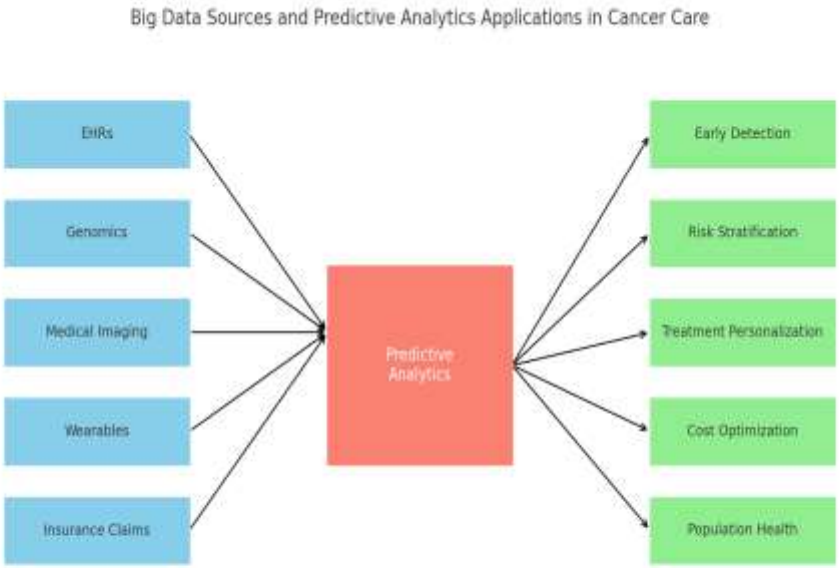

Figure 2. Big data sources and predictive analytics applications in oncology

Genomic and molecular datasets are another important domain. Advances in next-generation sequencing (NGS) have enabled large-scale analysis of somatic mutations, gene expression, epigenetic markers, and proteomics (Collins & Varmus, 2015; Garraway, 2013). These high-dimensional databases facilitate the identification of predictive biomarkers for cancer susceptibility, treatment response, and recurrence risk. Multi-omics integration, which combines genomes, transcriptomics, metabolomics, and microbiomics, improves prediction models by capturing the complexities of tumor biology (Hasin et al., 2017).

Medical imaging is another important source, as modalities like mammography, CT, MRI, and PET generate massive volumes of pixel-level data. Radiomics, or the computational extraction of quantitative features from images, captures tumor heterogeneity beyond human visual perception (Gillies et al., 2016). When combined with ML algorithms, radiomics has demonstrated potential in predicting tumor aggressiveness, treatment response, and survival outcomes (Aerts et al., 2014).

Claims and registration data provide insights into population trends and healthcare consumption patterns. Insurer claims data provide insights into cost trends, comorbidities, and treatment adherence (Miksad & Abernethy, 2018). Cancer registries, such as SEER in the United States, provide high-quality population-based data that improves model generalizability across demographics and geographic regions (Howlader et al., 2020).

Finally, wearables, cellphones, and remote monitoring devices collect real-time data on vital signs, activity levels, and medication adherence (Wang et al., 2018). These data streams allow for dynamic risk categorization and early detection of problems, especially for cancer survivors enduring long-term surveillance.

3.2 Analytical Techniques

The selection of analytical methods is contingent upon the characteristics of the data, the clinical inquiry, and the intended results.

Classical statistical models, including logistic regression, Cox proportional hazards models, and Bayesian inference, are extensively utilized for risk prediction because of their interpretability and seamless incorporation into clinical workflows (Harrell, 2015). These models excel with structured datasets of moderate dimensionality. Machine learning (ML) techniques demonstrate enhanced efficacy in managing high-dimensional, non-linear, and unstructured data. Supervised machine learning techniques,

including random forests, gradient boosting machines, and support vector machines (SVMs), have been utilized for cancer risk stratification, imaging categorization, and survival prediction (Kourou et al., 2015; Cruz & Wishart, 2006).

Deep learning (DL), especially convolutional neural networks (CNNs) and recurrent neural networks (RNNs), has transformed medical imaging and genomics by facilitating automatic feature extraction from raw data (Esteva et al., 2019; McKinney et al., 2020). Convolutional neural networks (CNNs) are proficient in radiology applications, although autoencoders and generative adversarial networks (GANs) have been utilized to enhance datasets and bolster generalization (Shen et al., 2017).

Natural language processing (NLP) enables the extraction of significant variables from unstructured clinical notes, pathology reports, and radiology narratives (Jensen et al., 2012). Transformer-based models such as BERT and BioBERT have enhanced the ability to analyze medical language for prediction applications (Lee et al., 2020).

Ensemble modeling, which integrates predictions from many techniques, improves robustness and predictive accuracy, especially in diverse datasets (Dietterich, 2000).

Emerging horizons encompass federated learning, which facilitates collaborative model training among institutions without direct data exchange, thereby safeguarding patient privacy and enhancing model generalizability (Li et al., 2020).

3.3 Model Validation and Interpretability

Validation is an important aspect of predictive analytics. Internal validation techniques like cross-validation and bootstrapping measure model performance within the training dataset, whereas external validation on different cohorts assesses generalizability (Steyerberg et al., 2019). Common performance measurements include area under the receiver operating characteristic curve (AUC), sensitivity, specificity, positive predictive value (PPV), and calibration (Harrell, 2015).

However, prediction accuracy is insufficient for clinical application. Interpretability is critical for developing clinician trust and providing meaningful insights. Traditional models, such as logistic regression, are easily interpreted, but complicated ML and DL models are sometimes criticized as "black boxes." SHAP (SHapley Additive Explanations), LIME (Local Interpretable Model-agnostic Explanations), and attention mechanisms in neural networks provide insight into model predictions (Ribeiro et al., 2016; Lundberg & Lee, 2017). Interpretable models are especially crucial in cancer due to the high stakes involved in diagnostic and therapy decisions.

3.4 Integration into Clinical Workflows

Integrating predictive analytics into healthcare procedures is critical for moving from research to practice. Clinical decision support systems (CDSS) incorporated in EHRs offer clinicians real-time risk assessments and recommendations at the point of care (Wright et al., 2017). User-friendly interfaces, interoperability with existing systems, and adherence to clinical recommendations are all necessary for successful integration.

Furthermore, implementation science highlights the value of multidisciplinary collaboration among oncologists, data scientists, informaticians, and policymakers in ensuring that predictive models are both technically sound and clinically relevant (Bates et al., 2014). Pilot studies and pragmatic trials are required to assess the real-world impact on outcomes and costs before general adoption (Neumann et al., 2018).

3.5 Ethical and Regulatory Considerations

Methodological rigor must be accompanied by compliance with ethical and regulatory requirements. Data privacy frameworks, such HIPAA in the United States and GDPR in Europe, provide limitations on data utilization and dissemination (Price & Cohen, 2019). Ethical issues encompass possible algorithmic prejudice, disparate access to predictive instruments, and the danger of excessive dependence on automated systems (Obermeyer et al., 2019). Consequently, governance mechanisms that guarantee openness, accountability, and equity are essential to scientific approaches in predictive oncology.

3.6 Summary

The methodological framework for utilizing big data and predictive analytics in cancer care is inherently interdisciplinary, necessitating the integration of varied datasets, advanced analytic techniques, stringent validation, interpretability tools, and meticulous clinical integration. By integrating methodological rigor with ethical and practical considerations, predictive analytics can be seen not just as a technical advancement but also as a transformative instrument for early identification and cost optimization in cancer.

4. Applications in Early Detection

Early cancer detection represents a significant opportunity for enhancing survival rates and decreasing treatment expenses. Predictive analytics, driven by big data, has facilitated a transition from conventional population-level screening to tailored, risk-adjusted, and data-informed early detection methodologies. Integrating multiple datasets, including imaging, pathology, genomes, and liquid biopsies, enables prediction models to detect cancers at an earlier stage, when therapies are less invasive, more successful, and more economical (Etzioni et al., 2003; Topol, 2019). This section examines applications across principal cancer types and underscores the revolutionary potential of multi-omics methodologies.

4.1 Breast Cancer

Breast cancer has long been a model disease for early identification via mammography, which has dramatically lowered mortality rates in high-income nations (Smith et al., 2019; Duffy et al., 2017). However, problems such as false positives, overdiagnosis, and low accuracy in dense breast tissue highlight the need for more advanced predictive techniques.

Artificial intelligence (AI) in mammography has made significant progress. Deep learning systems trained on millions of images have attained accuracy equal to or greater than that of expert radiologists. Google Health's deep learning system, for example, reduced false positives by 5.7% and false negatives by 9.4% in breast cancer screening when compared to radiologists in diverse populations (McKinney et al., 2020). Predictive models also support risk-adapted screening by incorporating polygenic risk scores (PRS), family history, and lifestyle factors to tailor screening intervals (Mavaddat et al., 2019).

Emerging modalities such as digital breast tomosynthesis (DBT), radiomics, and contrast-enhanced spectral mammography produce high-dimensional data that, when evaluated using machine learning, improves lesion diagnosis and characterization (Tagliafico et al., 2019). Furthermore, incorporating genetic indicators, such as BRCA1/2 mutations, into prediction models improves risk classification and allows for more targeted monitoring tactics (Garcia-Closas et al., 2014).

4.2 Lung Cancer

Lung cancer is the largest cause of cancer death worldwide, owing mostly to late-stage diagnosis (Siegel et al., 2022). Low-dose computed tomography (LDCT) has been demonstrated to lower mortality in high-risk smokers, however it has substantial false-positive rates (National Lung Screening Trial Research Team, 2011). Predictive analytics improves the value of LDCT by refining risk classification and decreasing unnecessary follow-up.

Machine learning algorithms that incorporate smoking history, demographic data, genetic susceptibility, and imaging markers have improved the prediction of malignancy in pulmonary nodules (Raji et al. 2020). Radiomics paired with deep learning has been particularly successful in distinguishing benign from malignant nodules with high accuracy (Hawkins et al., 2016). Furthermore, liquid biopsies that detect circulating tumor DNA (ctDNA), when combined with predictive algorithms, show promise for diagnosing early-stage lung cancer, especially in patients who do not qualify for LDCT screening (Cohen et al., 2018; Wan et al., 2017).

Predictive cost models used to LDCT screening show that modifying eligibility criteria using ML-based risk ratings improves cost-effectiveness while optimizing mortality reduction (Meza et al., 2020). Thus, predictive analytics not only improves early detection but also optimises screening resources.

4.3 Colorectal Cancer

Colorectal cancer (CRC) is significantly avoidable by early detection; however, the utilization of colonoscopy and fecal immunochemical assays (FIT) is inadequate (Wolf et al., 2018). Predictive analytics provides innovative methods for identifying high-risk individuals and improving engagement.

Machine learning algorithms have been utilized on electronic health records and claims data to identify persons at heightened risk of colorectal cancer, facilitating targeted screening outreach (Tanaka et al., 2019). The use of polygenic risk scores alongside lifestyle factors and family history enhances the identification of patients who could gain from earlier or more frequent colonoscopies (Hsu et al., 2015).

Novel biomarkers evaluated using predictive models, such as stool-based DNA assays and microbiome signatures, offer minimally intrusive screening options. Metagenomic sequencing of gut microbiota, in conjunction with machine learning methods, has exhibited significant sensitivity and specificity for the early identification of colorectal cancer (Yu et al., 2017). Radiomics utilized in CT colonography enhances the distinction between benign and malignant polyps (Ichikawa et al., 2018).

Predictive algorithms that estimate the probability of adenoma recurrence post-polypectomy facilitate the customization of surveillance intervals, hence alleviating patient burden and minimizing superfluous expenses (Corley et al., 2014).

4.4 Cervical Cancer

Cervical cancer continues to be a significant cause of death in low- and middle-income nations, despite its great preventability with screening and immunization (Arbyn et al., 2020). Traditional Pap smear cytology has been augmented with HPV DNA testing; however, predictive analytics further improves detection and risk stratification.

Deep learning models utilized for digital cytology images have attained accuracy on par with cytopathologists in identifying high-grade lesions (Zhang et al., 2019). The incorporation of HPV genotyping, viral load, and host gene methylation data into prediction models facilitates the categorization of women at elevated risk for progression to invasive cancer (Lorincz et al., 2016). Significantly, predictive analytics facilitates resource distribution in resource-constrained environments. Machine learning models utilizing demographic and behavioral data can pinpoint people at elevated risk, so informing mobile screening initiatives and enhancing effectiveness (Wentzensen & Schiffman, 2016). These methodologies illustrate how big data might mitigate worldwide disparities in early cancer detection.

4.5 Prostate Cancer

Prostate cancer screening with prostate-specific antigen (PSA) testing is contentious due to the potential for overdiagnosis and overtreatment (Loeb et al., 2014). Predictive analytics offers a means for more accurate and personalized risk evaluation.

Machine learning algorithms that integrate PSA dynamics, familial history, and genetic markers beyond the predictive capability of PSA alone in identifying clinically relevant prostate cancer (Murtola et al., 2018). Deep learning analysis of MRI-based radiomics has enhanced the discrimination between indolent and aggressive prostate tumors, hence minimizing unnecessary biopsies (Zhang et al., 2020).

The amalgamation of polygenic risk scores with conventional clinical criteria has enhanced screening methodologies, facilitating earlier identification in males at elevated risk while reducing harm to low-risk individuals (Seibert et al., 2018). These predictive methodologies correspond with cost optimization by diminishing the frequency of superfluous diagnostic procedures.

4.6 Multi-Omics and Integrated Approaches

The most transformational application of predictive analytics in early detection is found in multi-omics integration. Integrating genomic, transcriptomic, epigenomic, proteomic, metabolomic, and microbiomic data, prediction models encapsulate the multifaceted intricacies of cancer biology (Hasin et al., 2017).

Integrative models that amalgamate ctDNA methylation patterns, proteomic markers, and imaging characteristics have demonstrated elevated sensitivity in the early detection of various malignancies (Cohen et al., 2018). AI-driven technologies, such as multi-cancer early detection (MCED) assays, utilize extensive datasets and machine learning to concurrently screen for many malignancies with a single blood test (Liu et al., 2020).

Integrative techniques enhance sensitivity and specificity while potentially lowering costs by merging various single-cancer screening assays into a cohesive, minimally intrusive platform. Despite being in their nascent stages, preliminary evidence highlights their capacity to transform cancer screening methodologies.

4.7 Summary

Big data and predictive analytics applications in early cancer detection show promise for revolutionizing a variety of disease types. Predictive models improve diagnostic precision, customize screening, and maximize resource use in a variety of applications, including colorectal microbiome signatures, prostate MRI radiomics, and imaging for breast and lung cancer. A future where integrated AI-driven techniques allow for earlier, less expensive cancer detection at the population level is indicated by emerging multi-omics and MCED platforms. But achieving these potential calls for thorough verification, moral protections, and incorporation into just healthcare systems.

5. Cost Optimization in Cancer Care

The rising expenses of cancer treatment pose a significant issue for global healthcare systems. In 2020, cancer expenditures in the United States surpassed $200 billion and are anticipated to increase further due to an aging population, enhanced survivorship, and the implementation of innovative—frequently expensive—therapies such as immunotherapies and targeted medications (Mariotto et al., 2020; Yabroff et al., 2021). In low- and middle-income countries (LMICs), fundamental diagnostic and treatment infrastructure is insufficient, and resource limitations heighten the necessity for cost-effective care models

(Ginsburg et al., 2017). Predictive analytics offers a methodical strategy to enhance cost efficiency by forecasting patient requirements, mitigating inefficiencies, decreasing superfluous interventions, and aligning clinical pathways with value-based care tenets.

5.1 Reducing Unnecessary Diagnostics and Interventions

The excessive use of diagnostic and therapeutic interventions is a primary factor contributing to cancer-related expenses. The prevalent application of sophisticated imaging in asymptomatic individuals frequently results in incidental findings, prompting a series of superfluous tests, biopsies, and procedures (Welch & Black, 2010). Predictive analytics addresses this problem by facilitating risk-stratified diagnostic pathways.

Machine learning (ML) models that incorporate demographic, clinical, and genomic data can ascertain which patients are most predisposed to benefit from particular diagnostics. For instance, machine learning-based stratification has diminished superfluous biopsies in prostate cancer by differentiating indolent tumors from aggressive ones by MRI radiomics and PSA dynamics (Zhang et al., 2020; Seibert et al., 2018). Predictive modeling in breast cancer screening customizes mammography frequency based on individual risk, preventing over-screening in low-risk groups and concentrating resources on high-risk individuals (Mavaddat et al., 2019).

Predictive analytics also decreases therapeutic overtreatment. In early-stage breast and prostate cancers, recurrence risk models inform decisions regarding the necessity of aggressive chemotherapy versus active surveillance, enhancing quality of life and mitigating the expenses associated with superfluous treatment (Paik et al., 2004; Murtola et al., 2018).

5.2 Predicting and Preventing Hospital Readmissions

Hospital readmissions are a major cost burden in oncology. Unexpected readmissions within 30 days are prevalent due to complications such as infection, neutropenia, or severe medication reactions (Rathore et al., 2017). Predictive models based on EHR data, lab results, and prior admissions can effectively identify patients at high risk of readmission, allowing for proactive interventions such improved outpatient monitoring, early follow-up, or home-based care (Ong et al., 2017; Yu et al., 2018).

For example, in hematological malignancies, machine learning models predicted infection-related readmissions with great sensitivity, leading targeted preventative treatments and lowering inpatient care costs (Wang et al., 2019). Similarly, predictive analytics in palliative care identifies patients who are likely to undergo acute deterioration, allowing for early integration of home hospice services and avoiding costly emergency admissions (Smith et al., 2014).

5.3 Optimizing Treatment Selection and Resource Allocation

Cancer treatments, such as immunotherapies, targeted medicines, and precision oncology protocols, are some of the most costly elements of healthcare. Predictive analytics enhances therapy allocation by aligning patients with regimens that are most likely to provide benefits, therefore circumventing inefficient therapies and their related expenses.

Machine learning methods that incorporate genomic biomarkers, tumor mutational load, and immune microenvironment characteristics can predict patient responses to immune checkpoint inhibitors (Topalian et al., 2015; Kourou et al., 2015). These models mitigate the cost and clinical burden of non-response by directing treatment selection.

Predictive modeling additionally guides resource distribution within the healthcare system. Cost-prediction models estimate future expenses for cancer populations by analyzing incidence patterns, treatment usage, and survivability requirements (Miksad & Abernethy, 2018). These projections assist governments in budget allocation, workforce planning, and investment in efficient interventions like screening and preventative programs.

5.4 Value-Based Care and Predictive Analytics

The transition to value-based oncology treatment highlights outcomes attained in relation to expenses incurred. Predictive analytics is essential for implementing this transition. Stratifying patients based on clinical and financial risk enables prediction models to facilitate bundled payment arrangements, accountable care organizations (ACOs), and cancer care models (Porter, 2010; Neumann et al., 2018).

Clinical decision support systems (CDSS) that integrate predictive cost models assist oncologists in evaluating the trade-offs of treatment efficacy, side-effect profiles, and financial implications. Integrating cost-effectiveness data into treatment recommendations allows oncologists to discuss clinical and financial results with patients, promoting collaborative decision-making (Spratt et al., 2018).

At the payer level, predictive modeling enables the formulation of reimbursement frameworks that promote high-value treatment. Insurers are progressively utilizing predictive analytics to establish reimbursement rates for precision oncology medications based on anticipated outcomes and long-term cost savings (Sorenson et al., 2013). These frameworks guarantee reimbursement for costly therapies when substantiated by predictive evidence of efficacy.

5.5 Indirect Cost Reduction and Survivorship

In addition to direct healthcare costs, cancer incurs significant indirect expenses, such as productivity loss, long-term impairment, and caregiver strain (Bradley et al., 2011). Predictive workforce analytics has been employed to assess the economic ramifications of survivorship, guiding company policy for sick leave, workplace accommodations, and reintegration initiatives (Foysal Mahmud et al., 2024).

Predictive models of cancer survivorship trajectories identify individuals at risk of long-term incapacity, facilitating targeted vocational rehabilitation and mitigating production losses (Silver et al., 2013). Integrating social determinants of health into predictive models improves cost optimization by tackling issues including transportation constraints, food hardship, and housing instability, which lead to treatment non-adherence and increased downstream expenses (Marmot et al., 2020).

5.6 Global Implications

Cost optimization by predictive analytics is especially vital in low- and middle-income nations, where healthcare resources are few, and financial toxicity frequently hinders patients from finishing treatment. Predictive models can optimize the allocation of limited resources by prioritizing economical screening, customizing treatment intensity, and identifying populations most likely to benefit from subsidized treatments (Ginsburg et al., 2017). Predictive algorithms utilizing demographic and epidemiological data have been employed to distribute HPV vaccination and cervical cancer screening resources in sub-Saharan Africa, thereby optimizing population-level impact at lowest expense (Wentzensen & Schiffman, 2016).

5.7 Summary

Predictive analytics serves as a fundamental element for enhancing cancer care expenditures throughout the spectrum of prevention, diagnosis, treatment, and survivorship. Predictive models improve clinical and economical sustainability by minimizing superfluous tests and therapies, averting preventable readmissions, optimizing treatment distribution, and conforming to value-based frameworks. Moreover, by tackling secondary costs and directing global resource distribution, predictive analytics offers a comprehensive perspective on cancer economics. Despite ongoing problems in achieving equitable access, methodological transparency, and rigorous validation, the data highlights that cost optimization in cancer has evolved from a financial obligation to a clinical essential for sustainable healthcare provision.

6. Challenges and Ethical Considerations

The incorporation of big data and predictive analytics into cancer care presents significant potential, yet it also brings up a range of obstacles that must be meticulously managed to guarantee safe, equitable, and sustainable execution. The obstacles can be categorized into four main areas: data governance and privacy, interoperability and data quality, algorithmic bias and transparency, and legislative and equity issues.

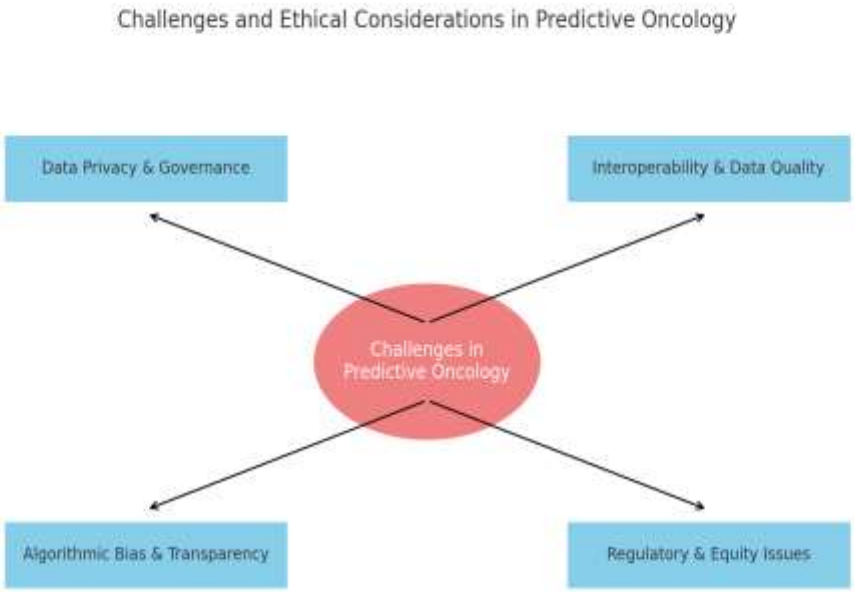

Figure 3. Challenges and ethical considerations in predictive oncology

6.1 Data Privacy and Governance

The delicate nature of oncology data, encompassing genetic sequences, imaging files, and comprehensive treatment histories, engenders substantial concerns regarding patient privacy and data security. Legal frameworks, such the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in Europe, impose stringent safeguards for personal health information (Price & Cohen, 2019). The intricacy of big data analytics frequently necessitates data exchange across institutions and borders, heightening the chance of breaches and unlawful utilization (Mittelstadt et al., 2016).

Innovative techniques like federated learning and privacy-preserving machine learning enable institutions to collaborate without the direct exchange of patient-level data, thus mitigating privacy hazards while ensuring analytical efficacy (Li et al., 2020). However, reconciling the necessity for extensive, representative datasets with the ethical obligation to safeguard individual privacy continues to be a significant difficulty.

6.2 Interoperability and Data Quality

Cancer data is derived from several sources—EHRs, genomics, imaging, wearables, and registries—that frequently employ incompatible formats and standards. The absence of interoperability obstructs the amalgamation of various datasets into unified predictive models (Murdoch & Detsky, 2013). Data quality concerns, including absent values, inaccurate coding, and unstructured text, diminish model dependability even inside EHR systems (Belle et al., 2015).

The Fast Healthcare Interoperability Resources (FHIR) standard has enhanced data interchange; yet, its global implementation is inconsistent (Mandel et al., 2016). In the absence of strong interoperability, predictive analytics may perpetuate silos instead of facilitating comprehensive, patient-centered insights. Investments in standardized ontologies, natural language processing (NLP) technologies, and quality-control pipelines are crucial for improving the applicability of big data in cancer.

6.3 Algorithmic Bias and Transparency

One of the most significant ethical concerns is algorithmic bias, which occurs when predictive algorithms reflect or exaggerate injustices seen in training data. Underrepresentation of minority groups in genomic datasets, for example, can result in predictive tools that perform poorly in non-European populations, compounding gaps in early detection and outcomes (Chen et al., 2020;

Obermeyer et al., 2019). Similarly, socioeconomic biases incorporated in EHRs can unfairly disfavor low-income or rural populations when employed in risk stratification models.

Transparency is another crucial concern. Many high-performing deep learning models are "black boxes," providing predictions without explicit explanations of the underlying factors (Ribeiro et al., 2016). In cancer, where clinical outcomes are critical, a lack of interpretability can weaken clinician trust and impede adoption. Tools like SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) improve transparency, but there is still a trade-off between model complexity and interpretability (Lundberg & Lee, 2017).

6.4 Regulatory and Equity Considerations

The high rate of innovation in predictive oncology frequently exceeds regulatory systems. The U.S. Food and Drug Administration (FDA) is still developing standards for AI-based medical devices, particularly those that learn and evolve on their own (FDA, 2021). Questions about culpability in the event of computational error remain unanswered, impeding clinical implementation.

Equity is another critical concern. Access to modern predictive analytics tools is unequal, with high-income countries and well-resourced institutions significantly more equipped to employ them than low- and middle-income countries (Kohli & Tan, 2016). Even within wealthy countries, gaps in access to genomic testing, high-quality imaging, and digital infrastructure threaten to exacerbate the cancer care gap (Marmot et al., 2020). Ethical implementation necessitates deliberate measures that promote fair access, such as subsidized programs, public-private partnerships, and capacity building in resource-constrained environments.

6.5 Summary

The incorporation of big data and predictive analytics into cancer care presents difficult issues that go beyond technological performance. To realize the full promise of predictive oncology, data privacy must be ensured, interoperability improved, algorithmic bias mitigated, regulatory paths clarified, and equity promoted. Addressing these challenges proactively is both an ethical and practical requirement for long-term adoption in a variety of healthcare settings.

7. Future Research Directions

As big data and predictive analytics continue to alter cancer care, future research must address present limits while broadening the spectrum of applications to assure greater effect. Several viable avenues emerge as critical for improving early detection and cost-effectiveness in cancer.

7.1 Federated Learning and Collaborative Analytics

One of the major issues in predictive oncology is data fragmentation across institutions and geographies. Federated learning provides a possible approach by allowing for collaborative model training without requiring direct exchange of sensitive patient-level data (Li et al., 2020). This strategy allows institutions to protect data privacy while contributing to larger, more representative models, which improves generalizability and reduces algorithmic bias. Future research should concentrate on creating safe, scalable federated learning frameworks for oncology that integrate genomes, imaging, and EHR data from across global populations.

7.2 Real-Time Predictive Analytics

The vast majority of current predictive models work retrospectively, evaluating past data to identify future dangers. However, real-time predictive analytics has the potential to change cancer care by delivering ongoing, dynamic risk assessments. Integration of streaming data from wearables, remote monitoring devices, and hospital systems may allow for proactive treatments such as early detection of adverse events, treatment toxicities, or disease progression (Wang et al., 2018). To enable real-time analytics at scale, researchers must overcome technical hurdles such as latency, processing efficiency, and clinical workflow integration.

7.3 Incorporating Social Determinants of Health (SDOH).

Cancer risk and outcomes are influenced not only by biological and clinical factors, but also by socioeconomic factors such as income, education, environment, and access to care (Marmot et al., 2020). Integrating SDOH into predictive models can improve accuracy, identify at-risk populations, and direct resource allocation techniques to reduce inequities. Future research should investigate standardized frameworks for collecting and incorporating SDOH data into predictive oncology, with a focus on ethical precautions against stigma and discrimination.

7.4 AI-Driven Clinical Trials and Adaptive Design

Traditional clinical trials are expensive, time-consuming, and frequently underestimate real-world populations. AI-powered clinical trials that use predictive analytics and adaptive designs could improve patient recruitment, stratify participants by risk, and alter procedures in real time depending on new data (Fleming et al., 2018). Predictive models could potentially increase trial efficiency by identifying biomarkers that are most likely to predict treatment response, resulting in faster medication development and lower costs. Future research should focus on verifying AI-driven trial frameworks under regulatory supervision in order to assure scientific rigor and patient safety.

Future predictive oncology research should focus on scalable, egalitarian, and real-world applications, rather than just proof-of-concept studies. Federated learning will allow for global collaboration, real-time analytics will assist proactive treatments, SDOH integration will alleviate disparities, and AI-powered trials will speed therapeutic innovation. Together, these paths point to a next-generation oncology ecosystem in which predictive analytics not only increase survival but also secures long-term and equitable cancer care around the world.

8. Conclusion

The incorporation of big data and predictive analytics in oncology signifies a transformative change in the detection, management, and funding of cancer. Utilizing varied datasets—such as electronic health records, genetics, imaging, registries, and patient-generated health data, predictive models can reveal trends that are undetectable by conventional diagnostic methods. This capability is especially revolutionary for early detection, since prompt diagnosis directly correlates with enhanced survival rates and diminished treatment intensity. Applications in breast, lung, colorectal, cervical, and prostate cancers illustrate how AI-enhanced imaging, genomics, radiomics, and liquid biopsies are transforming screening methodologies. Emerging multi-omics and multi-cancer early detection technologies enhance this potential by providing less invasive, integrative techniques that could transform population-level screening.

The importance of predictive analytics in cost optimization is equally significant, as cancer costs are increasingly rising worldwide. Predictive modeling enhances clinical efficiency and budgetary sustainability by minimizing superfluous tests and interventions, forecasting hospital readmissions, directing treatment choices, and facilitating value-based care frameworks. Indirect cost factors—such as survivability, productivity, and caregiver burden highlight the extensive economic importance of predictive oncology.

However, considerable obstacles and ethical dilemmas persist. Data privacy, interoperability, algorithmic bias, regulatory supervision, and equitable access must be systematically addressed to guarantee safe, transparent, and inclusive adoption. The risk of increasing gaps is especially pronounced in resource-constrained environments, when capacity enhancement and equitable policy structures are crucial.

Future research approaches indicate a focus on federated learning for global collaboration, real-time predictive analytics for dynamic care, the integration of social determinants of health to mitigate inequities, and AI-driven adaptive trials to expedite therapeutic innovation. Collectively, these advancements are poised to revolutionize oncology into a field that is not only more precise but also more proactive, equitable, and sustainable.

Utilizing big data and predictive analytics has transitioned from an aspirational objective to an imperative requirement, offering the twin potential of preserving lives and reducing expenses in the worldwide battle against cancer.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Aerts, H. J., Velazquez, E. R., Leijenaar, R. T., Parmar, C., Grossmann, P., Carvalho, S., … Lambin, P. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications, 5*(1), 4006. https://doi.org/10.1038/ncomms5006
[2] Arbyn, M., Weiderpass, E., Bruni, L., de Sanjosé, S., Saraiya, M., Ferlay, J., & Bray, F. (2020). Estimates of incidence and mortality of cervical cancer in 2018: A worldwide analysis. *The Lancet Global Health, 8*(2), e191–e203. https://doi.org/10.1016/S2214-109X(19)30482-6
[3] Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs, 33*(7), 1123–1131. https://doi.org/10.1377/hlthaff.2014.0041
[4] Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA, 319*(13), 1317–1318. https://doi.org/10.1001/jama.2017.18391

[5] Belle, A., Thiagarajan, R., Soroushmehr, S. M., Navidi, F., Beard, D. A., & Najarian, K. (2015). Big data analytics in healthcare. *BioMed Research International, 2015*, 370194. https://doi.org/10.1155/2015/370194

[6] Bradley, C. J., Yabroff, K. R., Dahman, B., Feuer, E. J., Mariotto, A., & Brown, M. L. (2011). Productivity costs of cancer mortality in the United States: 2000–2020. *Journal of the National Cancer Institute, 100*(24), 1763–1770. https://doi.org/10.1093/jnci/djn384

[7] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 71*(3), 209–249. https://doi.org/10.3322/caac.21660

[8] Chen, I. Y., Joshi, S., Ghassemi, M., & Topol, E. J. (2020). Treating health disparities with artificial intelligence. *Nature Medicine, 26*(1), 16–17. https://doi.org/10.1038/s41591-019-0649-2

[9] Chen, M., & Zhang, Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences, 275*, 314–347. https://doi.org/10.1016/j.ins.2014.01.015

[10] Cohen, J. D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., … Vogelstein, B. (2018). Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science, 359*(6378), 926–930. https://doi.org/10.1126/science.aar3247

[11] Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine, 372*(9), 793–795. https://doi.org/10.1056/NEJMp1500523

[12] Corley, D. A., Levin, T. R., Doubeni, C. A., Aden, A., Cheung, L. C., Zauber, A., … & Inadomi, J. (2014). Adenoma detection rate and risk of colorectal cancer and death. *New England Journal of Medicine, 370*(14), 1298–1306. https://doi.org/10.1056/NEJMoa1309086

[13] Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics, 2*, 59–77. https://doi.org/10.1177/117693510600200030

[14] Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems* (pp. 1–15). Springer.

[15] Duffy, S. W., Tabár, L., Yen, A. M., Dean, P. B., Smith, R. A., Jonsson, H., … Chen, T. H. (2017). Mammography screening reduces rates of advanced and fatal breast cancers: Results in 549,091 women. *Cancer, 124*(19), 3813–3822. https://doi.org/10.1002/cncr.31523

[16] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., … Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine, 25*(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

[17] Etzioni, R., Urban, N., Ramsey, S., McIntosh, M., Schwartz, S., Reid, B., … Anderson, G. (2003). The case for early detection. *Nature Reviews Cancer, 3*(4), 243–252. https://doi.org/10.1038/nrc1041

[18] FDA. (2021). Artificial intelligence and machine learning in software as a medical device. U.S. Food and Drug Administration. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

[19] Fleming, T. R., Labriola, D., & Wittes, J. (2018). Conducting clinical research during the COVID-19 pandemic: Protecting scientific integrity. *JAMA, 324*(1), 33–34. https://doi.org/10.1001/jama.2020.9286

[20] Garcia-Closas, M., Gunsoy, N. B., & Chatterjee, N. (2014). Combined associations of genetic and environmental risk factors: Implications for prevention of breast cancer. *Journal of the National Cancer Institute, 106*(11), dju305. https://doi.org/10.1093/jnci/dju305

[21] Garraway, L. A. (2013). Genomics-driven oncology: Framework for an emerging paradigm. *Journal of Clinical Oncology, 31*(15), 1806–1814. https://doi.org/10.1200/JCO.2012.46.8934

[22] Gillies, R. J., Kinahan, P. E., & Hricak, H. (2016). Radiomics: Images are more than pictures, they are data. *Radiology, 278*(2), 563–577. https://doi.org/10.1148/radiol.2015151169

[23] Ginsburg, O., Bray, F., Coleman, M. P., Vanderpuye, V., Eniu, A., Kotha, S. R., … & Conte, P. (2017). The global burden of women's cancers: A grand challenge in global health. *The Lancet, 389*(10071), 847–860. https://doi.org/10.1016/S0140-6736(16)31392-7

[24] Grosse, S. D., Teutsch, S. M., Haddix, A. C. (2008). Lessons from cost-effectiveness research for United States public health policy. *Annual Review of Public Health, 28*(1), 365–391. https://doi.org/10.1146/annurev.publhealth.28.021406.144058

[25] Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology, 18*, 83. https://doi.org/10.1186/s13059-017-1215-1

[26] Hawkins, S., Wang, H., Liu, Y., Garcia, A., Stringfield, O., Krewer, H., … Balagurunathan, Y. (2016). Predicting malignant nodules from screening CT scans. *Journal of Thoracic Oncology, 11*(12), 2120–2128. https://doi.org/10.1016/j.jtho.2016.07.002

[27] Hossain, S., … & Manik, M. M. T. G. (2024). Big Data Analysis and Prediction of COVID-19 Using Machine Learning Models in Healthcare. *Journal of Ecohumanism, 3*(8), 14468–. https://doi.org/10.62754/joe.v3i8.6775

[28] Howlader, N., Forjaz, G., Mooradian, M. J., Meza, R., Kong, C. Y., Cronin, K. A., … Mariotto, A. B. (2020). The effect of advances in lung-cancer treatment on population mortality. *New England Journal of Medicine, 383*(7), 640–649. https://doi.org/10.1056/NEJMoa1916623

[29] Hsu, L., Jeon, J., Brenner, H., Gruber, S. B., Schoen, R. E., Berndt, S. I., … & Newcomb, P. A. (2015). A model to determine colorectal cancer risk using common genetic susceptibility loci. *Gastroenterology, 148*(7), 1330–1339.e14. https://doi.org/10.1053/j.gastro.2015.02.010

[30] Ichikawa, D., Ishikawa, H., Kuroda, M., & Itoh, T. (2018). Machine learning for the prediction of colorectal cancer screening participation in Japan. *Applied Sciences, 8*(7), 1229. https://doi.org/10.3390/app8071229

[31] Inouye, M., Abraham, G., Nelson, C. P., Wood, A. M., Sweeting, M. J., Dudbridge, F., … & Samani, N. J. (2018). Genomic risk prediction of coronary artery disease in 480,000 adults. *Journal of the American College of Cardiology, 72*(16), 1883–1893. https://doi.org/10.1016/j.jacc.2018.07.079

[32] Institute of Medicine. (2013). *Delivering high-quality cancer care: Charting a new course for a system in crisis.* National Academies Press.

[33] Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics, 13*(6), 395–405. https://doi.org/10.1038/nrg3208

[34] Kohli, M., & Tan, S. Y. (2016). Electronic health records: How can IS help oncology care? *Cancer Journal, 22*(4), 263–266. https://doi.org/10.1097/PPO.0000000000000214

[35] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal, 13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

[36] Laney, D. (2001). 3D data management: Controlling data volume, velocity, and variety. *META Group Research Note*, 6, 70.

[37] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682
[38] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine, 37*(3), 50–60. https://doi.org/10.1109/MSP.2020.2975749
[39] Liu, M. C., Oxnard, G. R., Klein, E. A., Swanton, C., Seiden, M. V., & CCGA Consortium. (2020). Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Annals of Oncology, 31*(6), 745–759. https://doi.org/10.1016/j.annonc.2020.02.011
[40] Loeb, S., Bjurlin, M. A., Nicholson, J., Tammela, T. L., Penson, D. F., Carter, H. B., … & Catto, J. W. (2014). Overdiagnosis and overtreatment of prostate cancer. *European Urology, 65*(6), 1046–1055. https://doi.org/10.1016/j.eururo.2013.12.062
[41] Lorincz, A. T., Castanon, A., & Sasieni, P. (2016). New strategies for cervical cancer screening and prevention. *Current Opinion in Obstetrics and Gynecology, 28*(1), 4–9. https://doi.org/10.1097/GCO.0000000000000235
[42] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 30.
[43] Mandel, J. C., Kreda, D. A., Mandl, K. D., Kohane, I. S., & Ramoni, R. B. (2016). SMART on FHIR: A standards-based, interoperable apps platform for electronic health records. *Journal of the American Medical Informatics Association, 23*(5), 899–908. https://doi.org/10.1093/jamia/ocv189
[44] Manik, M. M. T. G. (2020). Biotech-Driven Innovation in Drug Discovery. *Journal of Computational Analysis and Applications, 28*(6), 41–47.
[45] Manik, M. M. T. G. (2021). Multi-Omics System with AI for Parkinson's Neurosurgery. *Journal of Medical and Health Studies, 2*(1), 42–52. https://doi.org/10.32996/jmhs.2021.2.1.5
[46] Manik, M. M. T. G. (2022). An Analysis of Cervical Cancer using the Application of AI and Machine Learning. *Journal of Medical and Health Studies, 3*(2), 67–76. https://doi.org/10.32996/jmhs.2022.3.2.11
[47] Manik, M. M. T. G. (2023). Multi-Omics Integration with ML for Early Detection of Ischemic Stroke. *Journal of Ecohumanism, 2*(2), 175–187. https://doi.org/10.62754/joe.v2i2.6800
[48] Manik, M. M. T. G., … (2020). The Role of Big Data in Combatting Antibiotic Resistance: Predictive Models for Global Surveillance. *Nanotechnology Perceptions, 16*(3), 361–378. https://doi.org/10.62441/nano-ntp.v16i3.5445
[49] Manik, M. M. T. G., … (2021). AI-Powered Predictive Analytics for Early Detection of Chronic Diseases. *Nanotechnology Perceptions, 17*(3), 269–288. https://doi.org/10.62441/nano-ntp.v17i3.5444
[50] Manik, M. M. T. G., Bhuiyan, M. M. R., Moniruzzaman, M., Islam, M. S., Hossain, S., & Hossain, S. (2018). The Future of Drug Discovery Utilizing Generative AI and Big Data Analytics for Accelerating Pharmaceutical Innovations. *Nanotechnology Perceptions, 14*(3), 120–135. https://doi.org/10.62441/nano-ntp.v14i3.4766
[51] Manik, M. M. T. G., Hossain, S., Ahmed, M. K., Rozario, E., Miah, M. A., Moniruzzaman, M., … Saimon, A. S. M. (2022). Integrating Genomic Data and Machine Learning to Advance Precision Oncology and Targeted Cancer Therapies. *Nanotechnology Perceptions, 18*(2), 219–243. https://doi.org/10.62441/nano-ntp.v18i2.5443
[52] Mariotto, A. B., Enewold, L., Zhao, J., Zeruto, C. A., & Yabroff, K. R. (2020). Medical care costs associated with cancer survivorship in the United States. *Cancer Epidemiology, Biomarkers & Prevention, 29*(7), 1304–1312. https://doi.org/10.1158/1055-9965.EPI-19-1534
[53] Marmot, M., Allen, J., Goldblatt, P., Herd, E., & Morrison, J. (2020). *Build back fairer: The COVID-19 Marmot review*. Institute of Health Equity.
[54] Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., … Easton, D. F. (2019). Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *American Journal of Human Genetics, 104*(1), 21–34. https://doi.org/10.1016/j.ajhg.2018.11.002
[55] McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., … Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature, 577*(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6
[56] Meza, R., Meernik, C., Jeon, J., & Cote, M. L. (2020). Lung cancer incidence trends by gender, race, and histology in the United States, 1973–2010. *PLoS ONE, 10*(3), e0121323. https://doi.org/10.1371/journal.pone.0121323
[57] Miah, M. A., Rozario, E., Khair, F. B., Ahmed, M. K., Bhuiyan, M. M. R., & Manik, M. M. T. G. (2019). Harnessing Wearable Health Data and Deep Learning Algorithms for Real-Time Cardiovascular Disease Monitoring and Prevention. *Nanotechnology Perceptions, 15*(3), 326–349. https://doi.org/10.62441/nano-ntp.v15i3.5278
[58] Miksad, R. A., & Abernethy, A. P. (2018). Harnessing big data to accelerate cancer research and improve patient care: The CancerLinQ initiative. *Journal of Clinical Oncology, 32*(34), 3727–3736. https://doi.org/10.1200/JCO.2014.56.7343
[59] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 2053951716679679. https://doi.org/10.1177/2053951716679679
[60] Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. *JAMA, 309*(13), 1351–1352. https://doi.org/10.1001/jama.2013.393
[61] Murtola, T. J., Tammela, T. L., Lahtela, J., Auvinen, A. (2018). Individualized screening for prostate cancer using PSA and polygenic risk scores. *European Urology, 73*(4), 551–556. https://doi.org/10.1016/j.eururo.2017.09.013
[62] Neumann, P. J., Cohen, J. T., & Weinstein, M. C. (2018). Updating cost-effectiveness—the curious resilience of the $50,000-per-QALY threshold. *New England Journal of Medicine, 371*(9), 796–797. https://doi.org/10.1056/NEJMp1405158
[63] Nilima, S. I. (2024). Advancement of Drug Discovery Using AI & ML. *IEEE COMPAS*. https://doi.org/10.1109/COMPAS60761.2024.10796748
[64] Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—Big data, machine learning,