Article published online: 2022-12-04

© 2022 IMIA and Georg Thieme Verlag KG

Clinical Informatics Approaches to Understand and Address Cancer Disparities

Tafadzwa L. Chaunzwa[1,2*], Maria Quiles del Rey[1*], Danielle S. Bitterman[1,2]

[1] Department of Radiation Oncology, Dana-Farber Brigham Cancer Center, Harvard Medical School, Boston, MA, USA

[2] Artificial Intelligence in Medicine (AIM) Program, Mass General Brigham, Harvard Medical School, Boston, MA, USA

* Contributed equally

Summary

Objectives: Disparities in cancer incidence and outcomes across race, ethnicity, gender, socioeconomic status, and geography are well-documented, but their etiologies are often poorly understood and multifactorial. Clinical informatics can provide tools to better understand and address these disparities by enabling high-throughput analysis of multiple types of data. Here, we review recent efforts in clinical informatics to study and measure disparities in cancer.

Methods: We carried out a narrative review of clinical informatics studies related to cancer disparities and bias published from 2018-2021, with a focus on domains such as real-world data (RWD) analysis, natural language processing (NLP), radiomics, genomics, proteomics, metabolomics, and metagenomics.

Results: Clinical informatics studies that investigated cancer disparities across race, ethnicity, gender, and age were identified. Most cancer disparities work within clinical informatics used RWD analysis, NLP, radiomics, and genomics. Emerging applications of clinical informatics to understand cancer disparities, including proteomics, metabolomics, and metagenomics, were less well represented in the literature but are promising future research avenues. Algorithmic bias was identified as an important consideration when developing and implementing cancer clinical informatics techniques, and efforts to address this bias were reviewed.

Conclusions: In recent years, clinical informatics has been used to probe a range of data sources to understand cancer disparities across different populations. As informatics tools become integrated into clinical decision-making, attention will need to be paid to ensure that algorithmic bias does not amplify existing disparities. In our increasingly interconnected medical systems, clinical informatics is poised to untap the full potential of multi-platform health data to address cancer disparities.

Keywords

Healthcare disparities; cancer; clinical informatics; big data; algorithms

Yearb Med Inform 2022:121-30

http://dx.doi.org/10.1055/s-0042-1742511

1 Introduction

Cancer is a leading cause of death worldwide [1, 2]. Approximately 10 million deaths in 2020 can be attributed to malignancies, most frequently carcinomas of the lung, colon, and liver [1, 2]. There are well known disparities in cancer incidence and outcomes across race, ethnicity, socioeconomic environment, sex, age, and geography. For example, breast cancer mortality among Black women in the United States (U.S.) is significantly higher than in other groups despite similar incidence rates [3]. Similarly, surveillance data from 2014 to 2019, as reported by the National Cancer Institute (NCI), show that Black men have the highest rate of new cancer diagnosis overall, while Asian/Pacific Islander men have the lowest [3-6].

Clinical informatics is playing an increasingly important role in decoding a wealth of multi-platform data to understand the complex interplay of social, economic, biologic, and environmental factors that contribute to cancer disparities. The integration of new high-throughput technology into scientific research is helping address important questions about the etiology as well as the genetic or molecular background of different cancers at a level not otherwise attainable with conventional methods. Here, we present a review of the key literature in the past two years exploring efforts to develop and implement informatics technologies to analyze these data and provide new insights on the determinants of cancer incidence and outcomes. We focus on clinical informatics domains such as real-world data (RWD) analysis, natural language processing (NLP), radiomics, genomics, proteomics, and metagenomics, which can be leveraged to better diagnose, treat, and understand cancer in diverse populations using a wide range of data streams. We also review studies on how bias may impact the interpretation and downstream effects of such efforts as they relate to disparities and discuss methods to identify and address bias.

2 Methods

For this narrative review, we performed a search of MEDLINE with a focus on prominent and frequently cited clinical informatics journals, including *JCO Clinical Cancer Informatics*, the *Journal of the American Medical Informatics Association*, and the *International Journal of Medical Informatics*. Articles that were published in 2018-2021 and relevant to our present discussion were reviewed, with emphasis on articles during the past two years and high impact articles from 2018 or 2019. Search terms including omics, radiomics, genomics, proteomics, metagenomics, disparities, and equity yielded a large number of publications which we narrowed to only those pertinent to oncology and alluding to differences defined by patients' racial or ethnic background, socioeconomic status, sex, age, geography, language/immigrant history, veteran status, and/or educational attainment.

3 Big Data and Real-World Data

Within a healthcare context, "big data" refers to large volumes of clinical information created by the adoption of digital technologies

and collected at one or more time points for large cohorts or individual patients [7]. RWD, which are data collected during routine patient care, are particularly valuable for timely, large-scale health outcomes research. The widespread adoption of electronic health records (EHRs) has facilitated collection and analysis of these data [8]. RWD have the potential to reveal and inform future mitigation of disparities because they provide an opportunity to characterize cancer outcomes among all patients receiving care, including groups often underrepresented in traditional prospective clinical trials and population studies [9].

A large retrospective prognostic cohort study by Peterson, et al., applied machine learning (ML) to identify patients with cancer receiving chemotherapy who were at increased risk for unplanned emergency department (ED) visits and hospital admissions [10]. A cohort of nearly 8,500 patients was analyzed using robust ML methods, including stratification of the test set by race, ethnicity, and insurance status. Black race and insurance by Medicaid, a U.S. national insurance for individuals with limited income, were found to be predictive of increased risk for preventable acute care use during chemotherapy, which may be associated with increased costs, worse outcomes, and negative overall patient experience. These findings are in line with prior observations suggesting Black, Hispanic, and Medicaid patients bear the brunt of cancer outcome inequities [11]. It is important to note, however, that inter-patient variability in data availability presents a challenge to the implementation of predictive models based on RWD in clinical oncology. In most real-world datasets, many patients lack recorded findings for important clinical factors (e.g., duration of therapy/follow up, data on long term outcomes, and baseline covariates). Even well-designed studies fall victim to missing data which can introduce bias and yield findings that may not be generalizable to vulnerable populations due to differences in healthcare access. For example, in the Peterson study above, 1,217/10,893 (11.2%) patients meeting the inclusion criteria were excluded because they were lost to follow up. There are well recognized challenges to the retention or continuation of care among non-White patients in the U.S. [12]. Notably, only 2.8% of patients in the 8,419-patient cohort analyzed by Peterson, et al., were Black. This is significantly lower than the proportion of individuals identifying as Black or African American (AA) in the U.S. general population (12.4%) [13]. The authors rightly pointed to the poor calibration of these data-driven ML models to underrepresented demographics, and the importance of making end-users of such clinical decision support tools aware of potential biases to mitigate the risk of perpetuating inequities. There are ongoing efforts to address these issues related to missing data. For example, Baron, et al., demonstrated the utility of an ensemble approach to predict patient-specific cancer survival and enable the construction of clinical predictive models that can accommodate interpatient heterogeneity in data availability [14].

There is growing interest in limiting preventable inequities in cancer care, however quantifying inequality is challenging. There are several relative and absolute measures for the quantification of healthcare disparities, and the optimal measure generally depends on context. Precise measurement of the magnitude of disparities and their temporal variation from RWD is critical and was the subject of a study assessing standard error estimation of confidence intervals for commonly used measures of health disparities in the literature [15]. This work evaluated the Health Disparities Calculator (HD*Calc), a free statistical software that calculates 11 commonly used health disparity measures and provides corresponding 95% confidence intervals (CIs) using either a Monte Carlo simulation-based method or an analytic method. Using age-adjusted cancer incidence rates from the NCI Surveillance, Epidemiology, and End Results Program (SEER) database [5] to conduct bias analyses, the authors concluded that HD*-Calc-generated CIs for some health disparity measures may be inaccurate in situations when data are sparse, such as in rare cancers or cancers where there is a large proportion of zero events across age group by social group combinations (a threshold of >25% was derived empirically). Accurate measurements of disparities could improve health equity by both identifying where disparities exist and facilitating social and economic risk-targeted care. For example, measures profiling risk based on social determinants of health, such as insurance status, language, and ethnicity, could be incorporated within EHRs to provide *in situ* clinical decision support for social risk-informed patient care [16].

Advances in electronic phenotyping are enabling scalable patient cohort creation and analysis to gain new insights into cancer disparities [17]. As an example, significant variability in metastatic breast cancer treatment and monitoring was observed across patient demographics and geographic region in a cohort of 6,180 U.S. women [18]. This cohort was identified via temporal data mining and the findings from the study suggest that, in addition to clinical factors, local resources and practice patterns influence individual treatment decisions. Similar clinical informatics tools have also been leveraged to assess guideline adherence in pediatric cancer cohorts. In a cohort of children exposed to chemotherapeutic agents that can cause cardiotoxicity, differences in guideline-based echocardiogram surveillance by sociodemographic factors such as race, ethnicity, and primary language were observed [19]. In this cohort, 87% of white patients received echocardiograms within the recommended time, compared with 76% of Black patients and only 55% of Hispanic patients. Regarding primary language, 90% of English-speaking patients compared with only 50% of Spanish-speaking patients received guideline-based care for echocardiogram surveillance. Further study is warranted to understand the root causes of these disparities and promote equitable survivorship-focused care in both pediatric and adult oncology [19]. These studies illustrate the promise of clinical informatics tools to generate cohorts for RWD analysis aimed at improving our understanding of cancer care disparities.

SEER and the NCI Cancer Research Data Commons are government initiatives that are helping eliminate data silos through harmonized data sharing and by providing access to large volumes of different data types. In the past year, analyses based on these large, collaborative data repositories have yielded insights into cancer disparities, as outlined here. While cancer screening programs are leading to better disease detection and im-

proved outcome, this improvement may not be shared by racial minorities and those with lower educational attainment [20, 21]. Race has also been shown to influence treatment recommendations, with Black patients less likely to be offered surgical resection for certain skull-base tumors [22]. Furthermore, excess cancer mortality has been reported in some U.S. minority groups across cancer types [23-32]. As an example, based on SEER analysis (2000-2017) the epidemiological profile of metastatic bladder cancer suggests Black females are more likely to die from this disease than any other group [33]. Across cancer types, targeted therapies are improving outcomes for patients with advanced disease. However, the promise of precision oncology continues to be elusive for individuals from low-socioeconomic backgrounds, who are less likely to undergo the requisite molecular profiling of their disease [34]. A similar trend is seen in the use of other specialized cancer treatments, such as brachytherapy for patients with cervical cancer [35, 36].

Novel computational methods for analyzing large volumes of data are also shedding new light into cancer inequities. Using a Naïve Bayesian network-based contribution analysis of biologic and clinical factors to cancer disparities, Luo, et al., found that nearly 50% of racial differences in stage at diagnosis for patients with breast cancer can be attributed to the timing and use of biopsy and screening mammography – modifiable and therefore actionable factors [37]. Additionally, a data matching algorithm was able to detect meaningful differences in the distribution of brain tumor histology between Veterans and non-Veterans populations, an approach that could be adapted to other sociodemographic factors [38, 39].

The COVID-19 pandemic has highlighted the power of informatics to rapidly analyze, synthesize, and act on RWD in near real-time. Marked racial and ethnic disparities in infections, COVID-19 deaths, and non-COVID-19 excess deaths have been observed since the start of the pandemic [40]. A large case-control study of deidentified EHR data from 73,449,510 patients across 360 hospitals in the U.S. found that patients with cancer were at a significantly higher risk of COVID-19 infection and severe disease [41]. Notably, Black patients were more likely to have COVID-19 than white patients, especially among patients with breast, prostate, colorectal, and lung cancer. The COVID-19 and Cancer Consortium (CCC19) is a multi-institutional registry of patients with COVID-19 and an existing or past cancer diagnosis. It has provided a unique opportunity to leverage RWD to understand the interactions between sociodemographic factors, a cancer diagnosis, and COVID-19 infection. Analyses of this cohort have found that race and ethnicity were not associated with mortality [42, 43], but that non-Hispanic Black race and Hispanic ethnicity were associated with more severe infection [44]. Further, Black patients with cancer in this registry were approximately half as likely to receive remdesivir as their white counterparts [43]. Such efforts demonstrate the power of clinical informatics to provide timely, high-quality evidence and illuminate health disparities.

4 Natural Language Processing

A major limitation of disparities research is that much of the clinical information, and especially information regarding race, ethnicity, and social determinants of health, has traditionally been documented as unstructured data in clinical text, and therefore is not readily analyzable at large scales. NLP, which aims to convert human language into representations that can be extracted and analyzed by computers, offers an avenue to glean the wealth of data within these texts to further our understanding of cancer care and outcomes across disparate populations [45-47]. Owing to major advances in deep learning algorithms for textual analysis, especially large contextual language models [48], NLP is now primed to make meaningful inroads in improving RWD analysis. There is an emerging body of work on cancer phenotyping and cohort development, but limited research into NLP methods to measure and assess cancer disparities [45-47]. One recent study used NLP to assist assessment of breast cancer guideline-concordant care from free text components of a cancer registry and found that receipt of non-guideline concordant care did not explain breast cancer mortality disparities across race [49]. Of note, the use of NLP to understand disparities is limited by the level of documentation of the sociodemographic factors. Agaronnik, *et al.,* developed an NLP pipeline to automate identification of patients with colorectal cancer and a chronic mobility disorder, a population with higher cancer-specific mortality, but results were limited by scarce documentation of patients' disabilities, highlighting a need for assessing and documenting these important disease-modifying factors [50].

NLP also has potential to reveal trends in medically underserved populations by mining and analyzing news and social media sources. One recent study used NLP, including sentiment analysis, to analyze web-based conversations about cancer clinical trials, and found that Black and Hispanic contributors had slightly more negative posts than white and Asian contributors. Differences in discussion of treatment stages and discussion topics were also identified, with Black contributors more likely to discuss costs and details of their healthcare professionals [51]. Such efforts reveal firsthand, patient-reported concerns that could underlie disparities in cancer care, and hint at the emerging value of medical-adjacent data to improve health equity.

In the future, NLP may help address disparities by automating resource-intensive processes that are currently disproportionately available in high socioeconomic settings. NLP-based clinical trial matching applications are being developed and may improve access to clinical trials in underrepresented populations [52, 53]. Similarly, NLP may also facilitate communication and self-management through patient- and provider-facing applications, which may improve healthcare access in traditionally underserved communities. For example, digital tools that integrate NLP to provide personalized screening and treatment recommendations based on social determinants of health have been proposed to facilitate broader access to personalized human papillomavirus vaccination and cancer screening recommendations [54].

5 Radiomics

Traditionally, clinical imaging studies have been qualitatively and subjectively interpreted by humans. Radiomics aims to quantitatively analyze and identify previously unrecognized patterns in images using high-throughput feature extraction. Relatedly, radiogenomics is defined as the linkage between radiographic phenotypes and genomic information [55]. In both cases, objective and precise quantitative imaging descriptors have the promise to serve as non-invasive prognostic or predictive biomarkers across cancer types and have demonstrated a capacity to capture intratumor heterogeneity and underlying gene-expression patterns [56]. While radiomics has previously relied on the explicit extraction of hand-crafted imaging features, more recent studies have shifted towards learned features obtained automatically from deep neural networks [57].

Age is a risk factor for cancer, and older individuals account for a large proportion of all patients with cancer. When compared to younger individuals, this population is more likely to be undertreated and excluded from clinical trials testing novel cancer therapeutics [58]. However, the older adult population is a heterogeneous group with significant variation in comorbidities and performance status. As such, chronological age may not fully capture cancer morbidity/mortality or accurately predict oncologic outcome [59]. Indeed, recent deep learning-based longitudinal multi-omics analyses have shown that chronological and biological age are not always concordant [60]. In light of this, better ways to quantify patients' true biological age are needed. Torres, *et al.,* used publicly available data from The Cancer Imaging Archive (TCIA) to construct and retrospectively validate a deep learning-based tool for lung cancer risk stratification [61]. They used pretreatment CT images to develop an "imaging-based prognostication technique" (IPRO) that performed mortality risk prediction in lung cancer with higher precision compared to TNM staging. In addition to risk stratification, another strength of the IPRO approach was that it was also able to effectively capture information regarding the biological age based on chest radiographs without being informed of the patient's chronological age.

Analyses based on hand-crafted/engineered radiomics features have also continued to shed new light into cancer racial disparities. A study comparing radiomics features in diverse populations with pancreatic ductal adenocarcinoma (PDAC) identified several textural radiomics features associated with unfavorable outcomes among Black patients with PDAC, independent of other prognostic factors such as tumor grade. The analytic dataset included cross sectional radiographs for 71 patients treated at a single institution [62].

A recent movement in "equitable machine learning" has stirred interest in studying the dangers of high-stakes AI-enabled predictive models that are used to inform practices in recruitment, law enforcement, and financial lending but are trained on unbalanced data [63-69]. Recent work in cancer radiomics has demonstrated a potential for similar ethical concerns in the medical domain due to imbalanced representation across populations. Studies with exciting results lacked demographic parity in their training and test sets which limits their generalizability to diverse populations based on race, sex, or other factors. A retrospective radiomics study by Wang, *et al.,* showed that ML-based analysis of magnetic resonance imaging (MRI) characteristics can predict tumor grade in soft tissue sarcomas. However, this study was limited by a small sample size that is not representative of the general sarcoma patient population, with 58 men and only 22 women in the training set [70]. Birra, *et al.,* introduced a novel outlier detection paradigm to better detect rare events using T1-weighted MRI radiomics features in glioblastoma. This approach differs from traditional binary classification in that it leverages class imbalance by modeling the non-outlier data objects [71]. It is important to note, however, that in this work a simple gaussian mixture model outperformed sophisticated deep learning frameworks, suggesting diminished utility of more complex solutions in small data settings. The authors highlighted this finding as it alludes to the pitfalls of blind reliance on ML, especially when input data is unbalanced. These challenges are not limited to imaging analysis, and equitable machine learning will require widespread recognition that improper application of ML to unbalanced datasets may lead to false conclusions that can ultimately amplify disparities.

6 Genomics

Genomics is an interdisciplinary field that studies gene abnormalities and gene expression networks driving the development and progression of tumors. Since 2006, when the first report of cancer genome sequencing appeared from The Cancer Genome Atlas Program (TCGA), key mutations have been found to be the molecular driver of different cancers, including in *BRCA1*, *TP53* and *RB1*.

Despite these successes, genome-wide association studies (GWAS), which are approaches used to associate specific genetic variations with disease, have been proven to not represent the broader population's genetic diversity, potentially exacerbating cancer disparities. The main contributing factor is that the samples used for genomic studies are often not representative of all genetic architectures [72, 73]. Therefore, population-specific variants may be missed, and the penetrance of the newly discovered genes and risk associations might not be accurately extrapolated to people with different ancestry. To address this issue, the Population Architecture using Genomics and Epidemiology study (PAGE) was created by the National Human Genome Research Institute. Based on PAGE data, a research group identified 27 novel loci and 38 ancestry-specific secondary signals at known loci, proving that with the appropriate sample size and mapping strategies, one can improve the discovery of complex genetic traits [74].

Likewise, the International Cancer Proteogenome Consortium (ICPC) aims to bring together 10 different countries to study the genetic and protein signatures of their most diagnosed cancers to unveil population-specific signatures. In addition, the New York Genome Center's Polyethnic-1000 Vision Program aims to study and include minorities' biological signature background into cancer treatment.

Efforts to improve the human reference genome have been made by using genetic sequencing information from 910 individuals of African descent. This work identified 296,485,284 base pairs, of which 387 fall within 315 distinct protein coding genes in the study population, identifying a set of

unique sequences that are specific for the African pan-genome and demonstrating that it contains 10% more DNA than the current human genome of reference [75].

In addition to efforts to increase the diversity of genomic information, there has been emerging work in identifying and addressing disparities via genomics studies. Davis, *et al.,* employed RNA sequencing to identify specific African ancestry genes that were upregulated in triple-negative breast cancer (TNBC), which disproportionately affects AA women. The studied population showed altered *TP53*, *NF1B* genes and AKT affected pathways, as well as down regulation of RNU2-6p. Furthermore, EGFR appeared to be a driver of residual TNBC in AA women [76]. Additionally, the prevalence of HER2+ breast cancer status in Latin American women is high, motivating an analysis of the genetic sequencing data from patients enrolled in the Peruvian Genetics and Genomics of Breast Cancer Study (PEGEN-BC). Their findings suggest that the odds of having a HER2+ breast cancer increased by a factor of 1.2 for every 10% increase in the Indigenous American ancestry, suggesting that the high prevalence of HER2+ breast cancer in Latin American women may be due to a specific genetic variant [77].

Mitochondria are organelles that are responsible for cellular energy metabolism, cell signaling, and oxidative stress. Dysregulation of this organelle is a hallmark of cancer. Mitochondrial genetic studies have been a focus of study to understand racial disparities in ovarian cancer. Changes in mtDNA-encoded genes, nuclear genes that encode for mitochondrial DNA, proteins within mitochondrial compartments, and molecular transporters may play a role in ovarian cancer disparity [78].

Compared to white men, Black men are 1.8 times more likely to be diagnosed with prostate cancer (PCa) and 2.2 times more likely to die from PCa [6]. A study analyzing next-generation sequencing (NGS) data from 205 samples of AA PCa patients showed that high percent of genome with copy-number alterations (PGA), somatic *TP53* mutations, and deletions in *CDNK1B* were associated with poor outcomes in AA men [79]. To investigate the relationship between the incidence/mortality and race/ethnic background in PCa, a group reviewed tumor genomic data from patients at two leading cancer centers. Four hundred seventy-four genes were studied across race (white, Black, Asian) and tumor stage (primary, metastatic). Among patients with primary PCa, druggable mutations were uncommon across the three groups. Among metastatic PCa patients, genes with existing targeted therapeutics, including DNA-repair genes and BRAF mutations, were more frequent in Black men than white men [80]. A comparative genomic study used targeted gene expression analysis on tumor mRNA to understand the different genetic pathways of PCa from West African men compared to AA men and white American men. This study found that prostate tumors in West African men have distinct genomic signatures, significant transcriptomic variability in androgen receptor-activity score, and are enriched for major proinflammatory pathways [81]. In another study of American men, AA men had upregulated expression of pathways related to immune response and increased response to DNA damage compared to European American men, who demonstrated increased expression of pathways related to DNA repair and WNT/beta-catenin signaling [82].

There is also a need for more diverse genomic data on non-small cell lung cancer (NSCLC), the leading cause of cancer death worldwide [1]. Genetic sequencing and analysis have revealed that AA individuals with lung squamous cell carcinoma have higher rates of chromothripsis and homologous recombination deficiency, which may lead to more aggressive tumor biology. Furthermore, they have higher frequencies of *PTEN* deletion and *KRAS* amplification [83]. A case control study revealed that lung adenocarcinomas of AA patients had significantly higher prevalence of mutated *PTPRT* and *JAK2*. Patients in the NCI-Md Case Control Study that had these mutations had increased IL-6/STAT3 signaling and miR21 expression [84]. Additionally, a study comparing the blood-based mutation profiles of Asian and white patients with NSCLC treated with atezolizumab found different *EFGR*, *TP53,* and *STK11* mutation profiles between the two groups [85].

Sex is another factor associated with cancer disparities. There is accumulating evidence that genes and proteins are differentially expressed between males and females. For example, genomics studies have linked sex with *p53* and *MDM2/4* mutations [86]. Another study that focuses on understanding why Kaposi sarcoma-associated herpesvirus preferentially infects and causes Kaposi sarcoma in males suggested that the androgen receptor is a functional prerequisite for cell invasion by Kaposi sarcoma-associated herpesvirus [87]. As part of the Pan/Cancer Analysis of Whole Genomes Consortium (PCAWG), Constance, *et al.,* reported an analysis of sex differences in whole genomes of 1,983 tumors, and found sex differences in non-coding autosomal genome, non-coding mutation density, tumor evolution, and mutation signatures [88].

Sex genetic differences may also influence treatment response. Ye, *et al.,* used large multi-omics data from TCGA to perform a comprehensive analysis of immune features across different cancer types to understand how immunotherapy efficacy may differ between male and female patients. They reported that male patients with melanoma had significantly higher tumor mutational burden (TMB), single nucleotide variation, neoantigen load, and PD-L1 expression [89]. Among patients with kidney renal papillary cell carcinoma, TMB, cytolytic activity, relative abundance of immune cells, and mRNA expression of immune checkpoints was higher among males compared to females. Female patients with lung squamous cell carcinomas exhibited higher levels of cytolytic activity and relative abundance of activated CD4 and CD8 +T cells, and had lower aneuploidy scores than males.

Females are at higher risk of developing papillary thyroid cancer (PTC). Han, *et al.*, identified 398 differentially expressed genes and 39 differentially expressed methylated genes between males and females with PTC, yielding new insights into sex differences in PTC [90].

As a note of caution, the existence of genetic variants does not necessarily explain cancer disparities. A study of Veterans Affairs (VA) patients with prostate cancer found that AA men did not present with later stage disease or have worse outcomes when

access to care is the same [91]. This study highlights that even if there are real genetic variants, they may not always drive differential outcomes. In fact, interpreting these differences as causal could stall progress in addressing disparities.

7 Emerging Applications

There are several exciting emerging applications of advanced informatics techniques to study and address cancer disparities. We highlight three areas of burgeoning research: proteomics, metabolomics, and metagenomics. Proteomics is the comprehensive characterization of proteins, their expression levels, patterns, interactions, and modifications within a cell or an organism. Similarly, metabolomics is the global study of small molecules or metabolites and can yield specific insight into cancer biology. There have been some recent studies focusing on the protein/metabolite differences that affect patient outcomes based on ancestry or sex background. AA men with PCa have been found to have higher exosome concentration levels when compared with healthy counterparts [92]. Proteomic analysis of the protein content in the exosomes found seven unique proteins in Black patients with PCa, and an increased inflammatory exosome content compared to healthy AA men. AA men with PCa protein content in the exosomes, when compared to healthy AA men, showed an upregulation of Filamin A. This protein was downregulated in exosomes of white men with PCa when compared to their race matched healthy control. Similarly, Ferrarini, *et al.*, identified distinct hepatocellular carcinoma metabolite signatures for AA, Asian, and white American patients [93]. These early efforts demonstrate the potential for proteomic and metabolomic analyses to offer new insights into the biological underpinning of observed cancer disparities.

Metagenomics is the study of the genetic material from a mixed microbial community, and recent studies have evaluated the impact of the microbiome on cancer [94]. While there is limited work on this topic in the timeframe of this review, noteworthy earlier findings suggest that the microbiota of patients with breast cancer is different from that of healthy controls, suggesting a possible role of microbes in the cancer environment with potential crosstalk between microbiota and endogenous hormones [95-97]. The microbiome can vary significantly between population groups, suggesting a possible role in cancer disparities. In a cohort of AA and white American patients with colorectal cancer, several microbes were differentially found in each group [98]. Further study of the microbiome could yield additional information on the drivers of cancer disparities.

8 Algorithmic Bias

Data-driven clinical prediction models have the potential to deliver clinical impact for the benefit of both patients and oncologists. Indeed, many oncology problems are ripe for AI applications [99-102]. However, algorithmic bias is a key and often underappreciated limitation to the clinical adoption of such methods [103-104]. When groups are underrepresented or have a skewed representation reflective of bias, racism, and inequities in the real world, models may be prone to systematic errors that disadvantage specific groups. The concordance between observed and predicted outcome is often low for subpopulations that are underrepresented in training sets of algorithms, resulting in reduced model performance for these groups [105]. The resulting bias can have the unintended consequence of propagating health disparities if such computational systems are not implemented with caution, and end-users need to be apprised of this potential danger as models enter clinical practice. Hendrix, *et al.*, studied primary care providers' (PCPs) preferences for AI technologies in breast cancer screening by asking how different model attributes impact the choice to recommend AI-enabled breast cancer screening. Among these attributes were sensitivity, specificity, radiologist involvement, understandability of AI decision-making, supporting evidence, and diversity of training data. Clinicians reported that an algorithm's sensitivity was more than twice as important as other attributes, including the diversity of the data used, in impacting their decision to recommend AI-enabled screening [106]. There could be several explanations for why clinicians chose other factors, such as sensitivity, over diversity of training data, including a need for more education on the potential harms and sources of bias in the clinical informatics era, and a historical emphasis on sensitivity as the most important metric for screening tests among clinicians. Another consideration in the setting of increasing clinical workloads is the unmet need for efficiency in the clinic, which may outweigh other considerations such as generalizability.

Bias in clinical prediction models can be improved via subgroup calibration [107]. Barda, *et al.*, studied the recalibration of predictions based on two common clinical prediction models using a multi-calibration fairness algorithm to protect against algorithmic discrimination [105, 107]. They evaluated predictions by the fracture risk assessment tool (FRAX) and Pooled Cohort Equations (PCE) on subgroups defined by ethnicity, socioeconomic status, age, sex, and immigrant status as well as in the overall population. The fairness algorithm was implemented for post-processing and significantly decreased the bias of subgroup mis-calibration, resulting in decreased algorithmic discrimination. While not focused on the oncology population, these findings could be extrapolated to similar clinical oncology tools. Other approaches to address bias include unsupervised ML with "biologic validation" of discoveries. Coombes, *et al.*, identified prognostic groups using unsupervised clustering of patients with chronic lymphocytic leukemia, a disease with well understood outcome determinants that could then provide biological validation for the prognostic groups [108]. These approaches may aid the cancer research community in realizing the goal of understanding and eliminating cancer disparities.

9 Conclusions

Clinical informatics will play an increasingly important role in efforts to narrow inequities in cancer care. A recent joint position statement by the American Association for Cancer Research (AACR), American Society of Clinical Oncology (ASCO), American

Cancer Society (ACS), and NCI emphasized the need to bring cutting-edge research tools to the study of cancer disparities including multi-omics platforms, as well as the need to engage the research community on how ancestry-informative markers can be integrated with sociodemographic data in oncology [109]. Clinical informatics is a promising avenue to decode the complex social and biological drivers of these disparities. Here, we presented a review of recent efforts to develop and use informatics applications for determining how demographic differences impact cancer outcomes. These and future efforts will be critical for the development of evidence-based strategies to mitigate inequities in cancer care.

It is important to recognize the potential dangers of large-scale informatics efforts aimed at investigating cancer disparities. Ethical concerns may arise given the risk of increasingly sophisticated clinical prediction models reflecting real human biases. Other less obvious ethical pitfalls have been discussed as computational models play a more prominent role in medicine [110]. Greater methodology reporting, including on details of missing data handling, will be paramount to ensuring that future informatics research serves to address, and not widen, disparities. Improved tracking of datasets through consensus identifiers and data linkage will also enhance transparency for published model evaluation in different populations, and facilitate evidence synthesis [111, 112].

The next phase of cancer disparities research will be driven by large-scale curation of multiple data streams in a multi-disciplinary setting [3, 113]. As we continue to collect finer-grained data on our patients, there will be enormous opportunities to apply informatics tools to improve cancer care for all. Collaborations between clinical oncologists, informaticians, public health officials, and, critically, researchers and representatives from the populations being studied, will be crucial for clinical informatics research to be translated into tangible improvements in cancer care and equity.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 2021 May;71(3):209-49.
2. Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, et al. Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer [Internet] 2020. [cited 2022 Jan 1]. Available from: https://gco.iarc.fr/today/home_
3. Hill HE, Schiemann WP, Varadan V. Understanding breast cancer disparities-a multi-scale challenge. Ann Transl Med 2020 Jul;8(14):906.
4. Hu S, Zhang W, Guo Q, Ye J, Zhang D, Zhang Y, et al. Prognosis and Survival Analysis of 922,317 Lung Cancer Patients from the US Based on the Most Recent Data from the SEER Database (April 15, 2021). Int J Gen Med 2021 Dec 10;14:9567-88.
5. Surveillance, Epidemiology, and End Results (SEER) Program Populations (1969-2019), National Cancer Institute, DCCPS, Surveillance Research Program [Internet] 2021 [cited 2022 Jan 1]. Available from: www.seer.cancer.gov/popdata
6. SEER*Explorer: An interactive website for SEER cancer statistics. Statistics. [Internet]. Surveillance Research Program, National Cancer Institute. [cited 2022 Apr 6]. Available from: https://seer.cancer.gov/explorer/
7. Pastorino R, De Vito C, Migliara G, Glocker K, Binenbaum I, Ricciardi W, et al. Benefits and challenges of Big Data in healthcare: an overview of the European initiatives. Eur J Public Health 2019 Oct 1;29(Supplement_3):23-7.
8. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-World Evidence - What Is It and What Can It Tell Us? N Engl J Med 2016 Dec 8;375(23):2293-7.
9. Lerman MH, Holmes B, St Hilaire D, Tran M, Rioth M, Subramanian V, et al. Validation of a Mortality Composite Score in the Real-World Setting: Overcoming Source-Specific Disparities and Biases. JCO Clin Cancer Inform 2021 Apr;5:401-13.
10. Peterson DJ, Ostberg NP, Blayney DW, Brooks JD, Hernandez-Boussard T. Machine Learning Applied to Electronic Health Records: Identification of Chemotherapy Patients at High Risk for Preventable Emergency Department Visits and Hospital Admissions. JCO Clin Cancer Inform 2021 Oct;5:1106-26.
11. Ward E, Jemal A, Cokkinides V, Singh GK, Cardinez C, Ghafoor A, et al. Cancer disparities by race/ethnicity and socioeconomic status. CA Cancer J Clin 2004 Mar-Apr;54(2):78-93.
12. Chhatre S, Malkowicz SB, Jayadevappa R. Continuity of care in acute survivorship phase, and short and long-term outcomes in prostate cancer patients. Prostate 2021 Dec;81(16):1310-9.
13. Jones N, Marks R, Ramirez R, Rios-Vargas M. 2020 Census Illuminates Racial and Ethnic Composition of the Country [Internet] 2021 [cited 2022 Jan 1]. Available from: https://www.census.gov/library/stories/2021/08/improved-race-ethnicity-measures-reveal-united-states-population-much-more-multiracial.html
14. Baron JM, Paranjape K, Love T, Sharma V, Heaney D, Prime M. Development of a "meta-model" to address missing data, predict patient-specific cancer survival and provide a foundation for clinical decision support. J Am Med Inform Assoc 2021 Mar 1;28(3):605-15.
15. Ahn J, Harper S, Yu M, Feuer EJ, Liu B, Luta G. Variance Estimation and Confidence Intervals for 11 Commonly Used Health Disparity Measures. JCO Clin Cancer Inform 2018 Dec;2:1-19.
16. Gold R, Sheppler C, Hessler D, Bunce A, Cottrell E, Yosuf N, et al. Using Electronic Health Record-Based Clinical Decision Support to Provide Social Risk-Informed Care in Community Health Centers: Protocol for the Design and Assessment of a Clinical Decision Support Tool. JMIR Res Protoc 2021 Oct 8;10(10):e31733.
17. Callahan A, Polony V, Posada JD, Banda JM, Gombar S, Shah NH. ACE: the Advanced Cohort Engine for searching longitudinal patient records. J Am Med Inform Assoc 2021 Jul 14;28(7):1468-79.
18. Caswell-Jin JL, Callahan A, Purington N, Han SS, Itakura H, John EM, et al. Treatment and Monitoring Variability in US Metastatic Breast Cancer Care. JCO Clin Cancer Inform 2021 May;5:600-14.
19. Noyd DH, Berkman A, Howell C, Power S, Kreissman SG, Landstrom AP, et al. Leveraging Clinical Informatics Tools to Extract Cumulative Anthracycline Exposure, Measure Cardiovascular Outcomes, and Assess Guideline Adherence for Children With Cancer. JCO Clin Cancer Inform 2021 Oct;5:1062-75.
20. Robinson TJ, Wilson LE, Marcom PK, Troester M, Lynch CF, Hernandez BY, et al. Analysis of Sociodemographic, Clinical, and Genomic Factors Associated With Breast Cancer Mortality in the Linked Surveillance, Epidemiology, and End Results and Medicare Database. JAMA Netw Open 2021 Oct 1;4(10):e2131020.
21. Ganesh A, Katipally R, Pasquinelli M, Feldman L, Spiotto M, Koshy M. Increased Disparities in Patients Diagnosed with Metastatic Lung Cancer Following Lung CT Screening in the United States. Clin Lung Cancer 2022 Mar;23(2):151-8.
22. Dhawan S, Alattar AA, Bartek J Jr, Ma J, Bydon M, Venteicher AS, et al. Racial disparity in recommendation for surgical resection of skull base chondrosarcomas: A Surveillance, Epidemiology, and End Results (SEER) analysis. J Clin Neurosci 2021 Dec;94:186-91.
23. Corona E, Yang L, Esrailian E, Ghassemi KA, Conklin JL, May FP. Trends in Esophageal Cancer Mortality and Stage at Diagnosis by Race and Ethnicity in the United States. Cancer Causes Control 2021 Aug;32(8):883-94.
24. Yan BY, Barilla S, Strunk A, Garg A. Survival differences in acral lentiginous melanoma according to socioeconomic status and race. J Am Acad Dermatol 2022 Feb;86(2):379-86.
25. Cheng JJ, Kim BJ, Kim C, Rodriguez de la Vega P, Varella M, Runowicz CD, Ruizet al. Association Between Race/Ethnicity and Survival in Women With Advanced Ovarian Cancer. Cureus 2021 Jun 30;13(6):e16070.
26. Long Parma D, Schmidt S, Muñoz E, Ramirez AG. Gastric adenocarcinoma burden and late-stage diagnosis in Latino and non-Latino populations in the United States and Texas, during

2004-2016: A multilevel analysis. Cancer Med 2021 Sep;10(18):6468-79.
27. Jawad MU, Bayne CO, Farhan S, Haffner MR, Carr-Ascher J, Alvarez E, et al. Prognostic factors, disparity, and equity variables impacting prognosis in bone sarcomas of the hand: SEER database review. J Surg Oncol 2021 Dec;124(8):1515-22.
28. Sempokuya T, Patel KP, Azawi M, Ma J, Wong LL. Increased morbidity and mortality of hepatocellular carcinoma patients in lower cost of living areas. World J Clin Cases 2021 Aug 16;9(23):6734-46.
29. Berkman AM, Andersen CR, Puthenpura V, Livingston JA, Ahmed S, Cuglievan B, et al. Disparities in the long-term survival of adolescent and young adult diffuse large B cell lymphoma survivors. Cancer Epidemiol 2021 Dec;75:102044.
30. Picon H, Guddati AK. Analysis of Trends in Mortality in Patients with Lymphoepithelial Carcinoma of the Head and Neck. Int J Gen Med 2021 Sep 29;14:6245-50.
31. Chiruvella V, Guddati AK. Analysis of Race and Gender Disparities in Mortality Trends from Patients Diagnosed with Nasopharyngeal, Oropharyngeal and Hypopharyngeal Cancer from 2000 to 2017. Int J Gen Med 2021 Oct 2;14:6315-23.
32. Lorona NC, Malone KE, Li CI. Racial/ethnic disparities in risk of breast cancer mortality by molecular subtype and stage at diagnosis. Breast Cancer Res Treat 2021 Dec;190(3):549-58.
33. Abdel-Rahman O. Bladder cancer mortality after a diagnosis of nonmuscle-invasive bladder carcinoma. Future Oncol 2019 Jul;15(19):2267-75.
34. Punekar SR, Griffin MM, Masri L, Roman SD, Makarov DV, Sherman SE, et al. Socioeconomic Determinants of the Use of Molecular Testing in Stage IV Colorectal Cancer. Am J Clin Oncol 2021 Dec 1;44(12):597-602.
35. Baig MZ, Filkins A, Khan M, Saif MW, Aziz H. Survival Benefits and Disparities in Adjuvant Radiation Therapy for Patients with Pancreatic Cancer. JOP 2021;22(2):36-41.
36. Boyce-Fappiano D, Nguyen KA, Gjyshi O, Manzar G, Abana CO, Klopp AH, Kamrava M, et al. Socioeconomic and Racial Determinants of Brachytherapy Utilization for Cervical Cancer: Concerns for Widening Disparities. JCO Oncol Pract 2021 Dec;17(12):e1958-e1967.
37. Luo Y, Carretta H, Lee I, LeBlanc G, Sinha D, Rust G. Naïve Bayesian network-based contribution analysis of tumor biology and healthcare factors to racial disparity in breast cancer stage-at-diagnosis. Health Inf Sci Syst 2021 Sep 24;9(1):35.
38. Woo C, Cioffi GN, Bej TA, Wilson B, Briggs JM, Markt SC, et al. Data Matching to Support Analysis of Cancer Epidemiology Among Veterans Compared With Non-Veteran Populations-An Exemplar in Brain Tumors. JCO Clin Cancer Inform 2021 Sep;5:985-94.
39. Alba PR, Gao A, Lee KM, Anglin-Foote T, Robison B, Katsoulakis E, et al. Ascertainment of Veterans With Metastatic Prostate Cancer in Electronic Health Records: Demonstrating the Case for Natural Language Processing. JCO Clin Cancer Inform 2021 Sep;5:1005-14.
40. Shiels MS, Haque AT, Haozous EA, Albert PS, Almeida JS, García-Closas M, et al. Racial and Ethnic Disparities in Excess Deaths During the COVID-19 Pandemic, March to December 2020. Ann Intern Med 2021 Dec;174(12):1693-9.
41. Wang Q, Berger NA, Xu R. Analyses of Risk, Racial Disparity, and Outcomes Among US Patients With Cancer and COVID-19 Infection. JAMA Oncol. 2021 Feb 1;7(2):220-7.
42. Kuderer NM, Choueiri TK, Shah DP, Shyr Y, Rubinstein SM, Rivera DR, et al; COVID-19 and Cancer Consortium. Clinical impact of COVID-19 on patients with cancer (CCC19): a cohort study. Lancet 2020 Jun 20;395(10241):1907-18.
43. Rivera DR, Peters S, Panagiotou OA, Shah DP, Kuderer NM, Hsu CY, et al; COVID-19 and Cancer Consortium. Utilization of COVID-19 Treatments and Clinical Outcomes among Patients with Cancer: A COVID-19 and Cancer Consortium (CCC19) Cohort Study. Cancer Discov 2020 Oct;10(10):1514-27.
44. Grivas P, Khaki AR, Wise-Draper TM, French B, Hennessy C, Hsu CY, et al. Association of clinical factors and recent anticancer therapy with COVID-19 severity among patients with cancer: a report from the COVID-19 and Cancer Consortium. Ann Oncol 2021 Jun;32(6):787-800.
45. Yim WW, Yetisgen M, Harris WP, Kwan SW. Natural Language Processing in Oncology: A Review. JAMA Oncol 2016 Jun 1;2(6):797-804.
46. Bitterman DS, Miller TA, Mak RH, Savova GK. Clinical Natural Language Processing for Radiation Oncology: A Review and Practical Primer. Int J Radiat Oncol Biol Phys 2021 Jul 1;110(3):641-55.
47. Savova GK, Danciu I, Alamudun F, Miller T, Lin C, Bitterman DS, Tourassi G, et al. Use of Natural Language Processing to Extract Clinical Cancer Phenotypes from Electronic Medical Records. Cancer Res 2019 Nov 1;79(21):5463-70.
48. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Internet]. arXiv [cs.CL]. 2018. Available from: http://arxiv.org/abs/1810.04805
49. Collin LJ, Yan M, Jiang R, Gogineni K, Subhedar P, Ward KC, et al. Receipt of Guideline-Concordant Care Does Not Explain Breast Cancer Mortality Disparities by Race in Metropolitan Atlanta. J Natl Compr Canc Netw 2021 Aug 16:jnccn20258.
50. Agaronnik ND, Lindvall C, El-Jawahri A, He W, Iezzoni LI. Challenges of Developing a Natural Language Processing Method With Electronic Health Records to Identify Persons With Chronic Mobility Disability. Arch Phys Med Rehabil 2020 Oct;101(10):1739-46.
51. Perez EA, Jaffee EM, Whyte J, Boyce CA, Carpten JD, Lozano G,. Analysis of Population Differences in Digital Conversations About Cancer Clinical Trials: Advanced Data Mining and Extraction Study. JMIR Cancer 2021 Sep 23;7(3):e25621.
52. Haddad T, Helgeson JM, Pomerleau KE, Preininger AM, Roebuck MC, Dankwa-Mullan I, et al. Accuracy of an Artificial Intelligence System for Cancer Clinical Trial Eligibility Screening: Retrospective Pilot Study. JMIR Med Inform 2021 Mar 26;9(3):e27767.
53. Beck JT, Rammage M, Jackson GP, Preininger AM, Dankwa-Mullan I, Roebuck MC, et al. Artificial Intelligence Tool for Optimizing Eligibility Screening for Clinical Trials in a Large Community Cancer Center. JCO Clin Cancer Inform 2020 Jan;4:50-9.
54. Olusanya OA, Ammar N, Davis RL, Bednarczyk RA, Shaban-Nejad A. A Digital Personal Health Library for Enabling Precision Health Promotion to Prevent Human Papilloma Virus-Associated Cancers. Front Digit Health 2021 Jul 21;3:683161.
55. Ak M, Toll SA, Hein KZ, Colen RR, Khatua S. Evolving Role and Translation of Radiomics and Radiogenomics in Adult and Pediatric Neuro-Oncology. AJNR Am J Neuroradiol 2022 Jun;43(6):792-801.
56. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014 Jun 3;5:4006. Erratum in: Nat Commun 2014;5:4644. Cavalho, Sara [corrected to Carvalho, Sara].
57. Chaunzwa TL, Hosny A, Xu Y, Shafer A, Diao N, Lanuti M, et al. Deep learning classification of lung cancer histology using CT images. Sci Rep 2021 Mar 9;11(1):5471.
58. Townsley CA, Selby R, Siu LL. Systematic review of barriers to the recruitment of older patients with cancer onto clinical trials. J Clin Oncol 2005 May 1;23(13):3112-24.
59. Schroyen S, Adam S, Jerusalem G, Missotten P. Ageism and its clinical impact in oncogeriatry: state of knowledge and therapeutic leads. Clin Interv Aging 2014 Dec 31;10:117-25.
60. Ahadi S, Zhou W, Schüssler-Fiorenza Rose SM, Sailani MR, Contrepois K, Avina M, et al. Personal aging markers and ageotypes revealed by deep longitudinal profiling. Nat Med 2020 Jan;26(1):83-90.
61. Torres FS, Akbar S, Raman S, Yasufuku K, Schmidt C, Hosny A, et al. End-to-End Non-Small-Cell Lung Cancer Prognostication Using Deep Learning Applied to Pretreatment Computed Tomography. JCO Clin Cancer Inform 2021 Oct;5:1141-50.
62. Permuth JB, Vyas S, Li J, Chen DT, Jeong D, Choi JW. Comparison of Radiomic Features in a Diverse Cohort of Patients With Pancreatic Ductal Adenocarcinomas. Front Oncol 2021 Jul 22;11:712950.
63. Wachter S, Mittelstadt B, Russell C. Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI (March 3, 2020). Computer Law & Security Review 41; 2021. Available from: http://dx.doi.org/10.2139/ssrn.3547922
64. Barrett LF, Adolphs R, Marsella S, Martinez AM, Pollak SD. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. Psychol Sci Public Interest 2019 Jul;20(1):1-68. Erratum in: Psychol Sci Public Interest 2019 Dec;20(3):165-6.
65. Furl N, Jonathon Phillips P, O'Toole AJ. Face

recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis. Cognitive Science [Internet]. 2002;26(6):797–815. Available from: http://dx.doi.org/10.1207/s15516709cog2606_4
66. Lahoti P, Gummadi KP, Weikum G. iFair: Learning Individually Fair Data Representations for Algorithmic Decision Making [Internet]. 2019 IEEE 35th International Conference on Data Engineering (ICDE). 2019. Available from: http://dx.doi.org/10.1109/icde.2019.00121
67. Buolamwini J, Friedler SA, Wilson C. Gender shades: Intersectional accuracy disparities in commercial gender classification [Internet]. [cited 2022 Apr 8]. Available from: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf
68. Barocas S, Selbst AD. Big Data's Disparate Impact [Internet]. 104 California Law Review 671. 2016. Available from: http://dx.doi.org/10.2139/ssrn.2477899
69. Eubanks, V. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Publishing Group; 2018.
70. Wang H, Chen H, Duan S, Hao D, Liu J. Radiomics and Machine Learning With Multiparametric Preoperative MRI May Accurately Predict the Histopathological Grades of Soft Tissue Sarcomas. J Magn Reson Imagin 2020 Mar;51(3):791-7.
71. Taha B, Li T, Boley D, Chen CC, Sun J. Detection of Isocitrate Dehydrogenase Mutated Glioblastomas Through Anomaly Detection Analytics. Neurosurgery 2021 Jul 15;89(2):323-8.
72. Davis MB. Genomics and Cancer Disparities: The Justice and Power of Inclusion. Cancer Discov 2021 Apr;11(4):805-9.
73. Zavala VA, Bracci PM, Carethers JM, Carvajal-Carmona L, Coggins NB, Cruz-Correa MR, et al. Cancer health disparities in racial/ethnic minorities in the United States. Br J Cancer 2021 Jan;124(2):315-32.
74. Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, et al. Genetic analyses of diverse populations improves discovery for complex traits. Nature 2019 Jun;570(7762):514-8.
75. Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat Genet 2019 Jan;51(1):30-5. Erratum in: Nat Genet 2019 Feb;51(2):364.
76. Davis M, Martini R, Newman L, Elemento O, White J, Verma A, et al. Identification of Distinct Heterogenic Subtypes and Molecular Signatures Associated with African Ancestry in Triple Negative Breast Cancer Using Quantified Genetic Ancestry Models in Admixed Race Populations. Cancers (Basel) 2020 May 13;12(5):1220.
77. Marker KM, Zavala VA, Vidaurre T, Lott PC, Vásquez JN, Casavilca-Zambrano S, et al; COLUMBUS Consortium. Human Epidermal Growth Factor Receptor 2-Positive Breast Cancer Is Associated with Indigenous American Ancestry in Latin American Women. Cancer Res 2020 May 1;80(9):1893-901.
78. Shukla P, Singh KK. The mitochondrial landscape of ovarian cancer: emerging insights. Carcinogenesis 2021 May 28;42(5):663-71.
79. Faisal FA, Murali S, Kaur H, Vidotto T, Guedes LB, Salles DC, et al. CDKN1B Deletions are Associated with Metastasis in African American Men with Clinically Localized, Surgically Treated Prostate Cancer. Clin Cancer Res 2020 Jun 1;26(11):2595-602.
80. Mahal BA, Alshalalfa M, Kensler KH, Chowdhury-Paulino I, Kantoff P, Mucci LA, et al. Racial Differences in Genomic Profiling of Prostate Cancer. N Engl J Med 2020 Sep 10;383(11):1083-5.
81. Yamoah K, Asamoah FA, Abrahams AOD, Awasthi S, Mensah JE, Dhillon J, et al. Prostate tumors of native men from West Africa show biologically distinct pathways-A comparative genomic study. Prostate 2021 Dec;81(16):1402-10.
82. Rayford W, Beksac AT, Alger J, Alshalalfa M, Ahmed M, Khan I, et al. Comparative analysis of 1152 African-American and European-American men with prostate cancer identifies distinct genomic and immunological differences. Commun Biol 2021 Jun 3;4(1):670.
83. Sinha S, Mitchell KA, Zingone A, Bowman E, Sinha N, Schäffer AA, et al. Higher prevalence of homologous recombination deficiency in tumors from African Americans versus European Americans. Nat Cancer 2020 Jan;1(1):112-21.
84. Mitchell KA, Nichols N, Tang W, Walling J, Stevenson H, Pineda M, et al. Recurrent PTPRT/JAK2 mutations in lung adenocarcinoma among African Americans. Nat Commun 2019 Dec 16;10(1):5735. Erratum in: Nat Commun 2020 Jan 30;11(1):700.
85. Qian J, Nie W, Lu J, Zhang L, Zhang Y, Zhang B, et al. Racial differences in characteristics and prognoses between Asian and white patients with nonsmall cell lung cancer receiving atezolizumab: An ancillary analysis of the POPLAR and OAK studies. Int J Cancer 2020 Jun 1;146(11):3124-33.
86. Mancini F, Giorgini L, Teveroni E, Pontecorvi A, Moretti F. Role of Sex in the Therapeutic Targeting of p53 Circuitry. Front Oncol 2021 Jul 8;11:698946.
87. Ding M, Wu J, Sun R, Yan L, Bai L, Shi J, et al. Androgen receptor transactivates KSHV noncoding RNA PAN to promote lytic replication-mediated oncogenesis: A mechanism of sex disparity in KS. PLoS Pathog 2021 Sep 20;17(9):e1009947.
88. Li CH, Prokopec SD, Sun RX, Yousif F, Schmitz N; PCAWG Tumour Subtypes and Clinical Translation, Boutros PC; PCAWG Consortium. Sex differences in oncogenic mutational processes. Nat Commun 2020 Aug 28;11(1):4330.
89. Ye Y, Jing Y, Li L, Mills GB, Diao L, Liu H, et al. Sex-associated molecular differences for cancer immunotherapy. Nat Commun 2020 Apr 14;11(1):1779.
90. Han R, Sun W, Huang J, Shao L, Zhang H. Sex-biased DNA methylation in papillary thyroid cancer. Biomark Med 2021 Feb;15(2):109-20.
91. Riviere P, Luterstein E, Kumar A, Vitzthum LK, Deka R, Sarkar RRet al. Survival of African American and non-Hispanic white men with prostate cancer in an equal-access health care system. Cancer 2020 Apr 15;126(8):1683-90.
92. Panigrahi GK, Praharaj PP, Kittaka H, Mridha AR, Black OM, Singh R, et al. Exosome proteomic analyses identify inflammatory phenotype and novel biomarkers in African American prostate cancer patients. Cancer Med 2019 Mar;8(3):1110-23.
93. Ferrarini A, Di Poto C, He S, Tu C, Varghese RS, Kara Balla A, et al. Metabolomic Analysis of Liver Tissues for Characterization of Hepatocellular Carcinoma. J Proteome Res 2019 Aug 2;18(8):3067-76.
94. Wei LQ, Cheong IH, Yang GH, Li XG, Kozlakidis Z, Ding L, et al. The Application of High-Throughput Technologies for the Study of Microbiome and Cancer. Front Genet 2021 Jul 28;12:699793.
95. Chen J, Douglass J, Prasath V, Neace M, Atrchian S, Manjili MH, et al. The microbiome and breast cancer: a review. Breast Cancer Res Treat 2019 Dec;178(3):493-6.
96. Chen D, Wu J, Jin D, Wang B, Cao H. Fecal microbiota transplantation in cancer management: Current status and perspectives. Int J Cancer 2019 Oct 15;145(8):2021-31.
97. Fernández MF, Reina-Pérez I, Astorga JM, Rodríguez-Carrillo A, Plaza-Díaz J, Fontana L. Breast Cancer and Its Relationship with the Microbiota. Int J Environ Res Public Health 2018 Aug 14;15(8):1747.
98. Farhana L, Antaki F, Murshed F, Mahmud H, Judd SL, Nangia-Makker P, et al. Gut microbiome profiling and colorectal cancer in African Americans and Caucasian Americans. World J Gastrointest Pathophysiol 2018 Sep 29;9(2):47-58.
99. Kann BH, Hosny A, Aerts HJW. Artificial intelligence for clinical oncology. Cancer Cell [Internet] 2021;39(7):916–27. Available from: http://dx.doi.org/10.1016/j.ccell.2021.04.002
100. Huynh E, Hosny A, Guthier C, Bitterman DS, Petit SF, Haas-Kogan DA, et al. Artificial intelligence in radiation oncology. Nat Rev Clin Oncol [Internet] 2020;17(12):771–81. Available from: http://dx.doi.org/10.1038/s41571-020-0417-8
101. Lu MT, Raghu VK, Mayrhofer T, Aerts HJWL, Hoffmann U. Deep Learning Using Chest Radiographs to Identify High-Risk Smokers for Lung Cancer Screening Computed Tomography: Development and Validation of a Prediction Model. Ann Intern Med 2020 Nov 3;173(9):704-13.
102. Xu Z, Wang X, Zeng S, Ren X, Yan Y, Gong Z. Applying artificial intelligence for cancer immunotherapy. Acta Pharm Sin B 2021 Nov;11(11):3393-405.
103. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019 Oct 29;17(1):195.
104. Singh RP, Hom GL, Abramoff MD, Campbell JP, Chiang MF; AAO Task Force on Artificial Intelligence. Current Challenges and Barriers to Real-World Artificial Intelligence Adoption for the Healthcare System, Provider, and the Patient. Transl Vis Sci Technol 2020 Aug 11;9(2):45.
105. Barda N, Yona G, Rothblum GN, Greenland P, Leibowitz M, Balicer R, et al. Addressing bias in prediction models by improving subpopulation calibration. J Am Med Inform Assoc 2021 Mar

1;28(3):549-58.
106. Hendrix N, Hauber B, Lee CI, Bansal A, Veenstra DL. Artificial intelligence in breast cancer screening: primary care provider preferences. J Am Med Inform Assoc 2021 Jun 12;28(6):1117-24.
107. Hébert-Johnson Ú, Kim MP, Reingold O, Rothblum GN. Calibration for the (Computationally-Identifiable) Masses [Internet]. arXiv [cs.LG] 2017. Available from: http://arxiv.org/abs/1711.08513
108. Coombes CE, Abrams ZB, Li S, Abruzzo LV, Coombes KR. Unsupervised machine learning and prognostic factors of survival in chronic lymphocytic leukemia. J Am Med Inform Assoc 2020 Jul 1;27(7):1019-27.
109. Polite BN, Adams-Campbell LL, Brawley OW, Bickell N, Carethers JM, Flowers CR, et al. Charting the Future of Cancer Health Disparities Research: A Position Statement From the American Association for Cancer Research, the American Cancer Society, the American Society of Clinical Oncology, and the National Cancer Institute. J Clin Oncol 2017 Sep 10;35(26):3075-82.
110. Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. N Engl J Med 2018 Mar 15;378(11):981-3.
111. Crowley RJ, Tan YJ, Ioannidis JPA. Empirical assessment of bias in machine learning diagnostic test accuracy studies. J Am Med Inform Assoc 2020 Jul 1;27(7):1092-101.
112. Block RG, Puro J, Cottrell E, Lunn MR, Dunne MJ, Quiñones AR, et al. Recommendations for improving national clinical datasets for health equity research. J Am Med Inform Assoc 2020 Nov 1;27(11):1802-7.
113. Ma C, Sridharan M, Al-Sayegh H, Li A, Guo D, Auclair M, et al. Building a Harmonized Datamart by Integrating Cross-Institutional Systems of Clinical, Outcome, and Genomic Data: The Pediatric Patient Informatics Platform (PPIP). JCO Clin Cancer Inform 2021 Feb;5:202-15.

Correspondence to:
Dr. Danielle S. Bitterman
Department of Radiation Oncology
Dana-Farber Cancer Institute/Brigham and Women's Hospital
75 Francis Street
Boston, MA 02115, USA
Tel: +1 857 215 1489
Fax: +1 617 975 0985
E-mail: Danielle_Bitterman@dfci.harvard.edu