Clinical Informatics Approaches to Understand and Address Cancer Disparities

Tafadzwa L. Chaunzwa1,2, Maria Quiles del Rey1, Danielle S. Bitterman1,2

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

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 rate 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-
improved 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 Naive 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 first-hand, 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].

IMIA Yearbook of Medical Informatics 2022
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 data-set 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 CDKN1B 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 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 PTPR1 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 EGFR, 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 present 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


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