Journal of the National Cancer Institute Monographs, 2023, **2023(62)**, 246–254
https://doi.org/10.1093/jncimonographs/lgad025
Monograph

Data gaps and opportunities for modeling cancer health equity

Amy Trentham-Dietz, PhD,[1,*] Douglas A. Corley, MD, PhD,[2] Natalie J. Del Vecchio, PhD,[3] Robert T. Greenlee, PhD, MPH,[4] Jennifer S. Haas, MD, MSc,[5] Rebecca A. Hubbard, PhD,[6] Amy E. Hughes, PhD,[7] Jane J. Kim, PhD,[8] Sarah Kobrin, PhD, MPH,[9] Christopher I. Li, MD, PhD,[3] Rafael Meza, PhD,[10] Christine M. Neslund-Dudas, PhD,[11] Jasmin A. Tiro, PhD, MPH[12]

[1]Department of Population Health Sciences and Carbone Cancer Center, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI, USA
[2]Division of Research, Kaiser Permanente Northern California, Oakland, CA, USA
[3]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[4]Marshfield Clinic Research Institute, Marshfield, WI, USA
[5]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA
[6]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[7]Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA
[8]Department of Health Policy and Management, Center for Health Decision Science, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[9]Healthcare Delivery Research Program, Division of Cancer Control & Population Sciences, National Cancer Institute, National Institutes of Health, Rockville, MD, USA
[10]Department of Integrative Oncology, British Columbia (BC) Cancer Research Institute, Vancouver, BC, Canada
[11]Department of Public Health Sciences and Henry Ford Cancer, Henry Ford Health, Detroit, MI, USA
[12]Department of Public Health Sciences, University of Chicago Biological Sciences Division, and University of Chicago Medicine Comprehensive Cancer Center, Chicago, IL, USA

*Correspondence to: Amy Trentham-Dietz, PhD, Department of Population Health Sciences and Carbone Cancer Center, School of Medicine and Public Health, University of Wisconsin-Madison, 610 Walnut St WARF Room 307, Madison, WI 53726, USA (e-mail: trentham@wisc.edu).

Abstract

Population models of cancer reflect the overall US population by drawing on numerous existing data resources for parameter inputs and calibration targets. Models require data inputs that are appropriately representative, collected in a harmonized manner, have minimal missing or inaccurate values, and reflect adequate sample sizes. Data resource priorities for population modeling to support cancer health equity include increasing the availability of data that 1) arise from uninsured and underinsured individuals and those traditionally not included in health-care delivery studies, 2) reflect relevant exposures for groups historically and intentionally excluded across the full cancer control continuum, 3) disaggregate categories (race, ethnicity, socioeconomic status, gender, sexual orientation, etc.) and their intersections that conceal important variation in health outcomes, 4) identify specific populations of interest in clinical databases whose health outcomes have been understudied, 5) enhance health records through expanded data elements and linkage with other data types (eg, patient surveys, provider and/or facility level information, neighborhood data), 6) decrease missing and misclassified data from historically underrecognized populations, and 7) capture potential measures or effects of systemic racism and corresponding intervenable targets for change.

For more than 20 years, the US National Cancer Institute (NCI) has supported the Cancer Intervention and Surveillance Modeling Network (CISNET) to quantify the impact of changes in risk factors [eg, smoking cessation (1,2), HPV vaccines (3)] and advances in screening and therapy on population cancer mortality over time (4-6). The CISNET models have also been used to project the impact of various hypothetical screening guidelines on the cancer burden (7-10). CISNET modeling teams have examined cancer outcomes for groups with elevated risk of cancer mortality, such as women with pathogenic mutations placing them at greater risk of breast cancer (11), and the risk of lung cancer mortality among adults who smoke tobacco cigarettes (2). CISNET has also conducted studies of population groups that experience cancer disparities, in other words, adverse differences in cancer prevention, incidence, stage at diagnosis, tumor subtype, and/or mortality among people who have been historically underrecognized in health care and face greater obstacles to health (12), although such studies are fewer in number. For example, the CISNET prostate model teams have quantified prostate survival disparities for Black and African American (hereafter Black) men as compared with men overall after accounting for overdiagnosis and lead time because of screening (13). Recent modeling reports have also evaluated whether existing mammography screening strategies are equitable for Black women (14), lung cancer screening disparities among Black adults (15), the potential effects of racism in colorectal cancer incidence and outcomes (16), and cervical cancer screening using self-sampling as an approach for reducing disparities among Black women in the Mississippi Delta (17).

Opportunities for researchers to use population modeling as a tool for designing strategies to improve health equity are enhanced, or conversely limited, by the quality of model

Received: April 29, 2023. **Revised:** July 12, 2023. **Accepted:** August 15, 2023

parameter inputs and calibration targets, which is influenced by the strength of the underlying data. CISNET models use the strongest nationally representative evidence provided by empirical studies and data resources as inputs and calibration targets. Multiple data inputs are incorporated in model structures and can consequently reflect how factors interact along the cancer control continuum, from risk and prevention (eg, smoking, body mass index) to diagnosis (eg, cancer stage depending on method of detection) to end of life (eg, quality of life) (18-20). Barriers to obtaining high-quality data inputs are numerous and vary across model components. Importantly, these barriers can be greater for at-risk populations including persons defined by one or more social categories (eg, race, ethnicity, gender, sexual orientation, socioeconomic status), which increases complexity for inclusion in population models (21). Elsewhere in this Monograph, Chapman and colleagues (12) have outlined a framework for conceptualizing these types of interactions, and we use this framework to guide our consideration of data gaps and opportunities to facilitate future modeling to test the effects of interventions on cancer equity. In this report, we also describe strengths and limitations of data currently used as inputs to various model components and identify data resource needs that would expand capability for modeling strategies to ultimately alleviate cancer disparities and achieve health equity with an emphasis on racial equity.

Strengths and limitations of model input data

Population models of cancer include multiple components along the cancer control continuum operating at different levels from the cellular to the population level, all of which need to be informed by data inputs or calibration targets (18-20). Specifying a proposed causal mechanism between the components of a model can be important for improving transparency and credibility. However, models do not necessarily have to specify the mechanistic pathways by which components are causally related, and modeling may not be the best tool for answering certain research questions where input data are unavailable or poor quality, such as why some patients experience delays in diagnosis or treatment initiation due to barriers to care. Such topics may best be pursued using other study designs involving primary data collection. Yet, models can provide insight into actionable steps to ameliorate disparities at different points along the cancer control continuum. For example, cancer mortality disparities can reflect the interplay of race and cancer subtypes (eg, lung cancer histologic type or colorectal cancer anatomic location) affected by residential segregation including exposure to predatory tobacco and alcohol marketing, environmental injustice (ie, industrial facility proximity), and limited access to healthy foods and health care, as well as reduced educational and occupational opportunities (Table 1, upstream factors). Models that include natural history components may explicitly include parameters that reflect the biological effects of adversity, in other words, the effect of chronic stressors on physiology and epigenetics (12). Modeling teams must carefully consider how models are structured and how input data are used singly and jointly within the model. For example, if tumor growth rates tend to be faster in one population group compared with another prior to diagnosis because of greater risk factor exposure, the population-specific distributions of stage at cancer diagnosis and their corresponding survival rates should correspond to higher mortality rates. Researchers face challenges to develop models that are well calibrated at all points along the cancer continuum, especially for modeling populations (eg, Black adults) with greater cancer burdens, because models may require more reprogramming, parameterization, and calibration rather than simple data input substitutions.

Health-care data used to inform model inputs may suffer from biases driven by inequities in care (Table 1, detection, diagnosis, treatment) or biases in patient selection for model inputs. For example, the 12% prevalence of a family history of breast cancer based on the Breast Cancer Surveillance Consortium (BCSC, a research network of academic and community-based breast imaging clinics) (22) is almost double that of the 7% self-reported prevalence observed in the National Health Interview Survey (a surveillance study recruiting participants through random sampling of households) (23). Some factors may be over- or underrepresented in the BCSC compared with the general population because of the geographic location of the BCSC registries or referral patterns for breast imaging patients in the BCSC with potentially elevated risk of breast cancer or because persons who never obtain a mammogram because of barriers to obtaining health care are not included in the BCSC. Data from household studies are also limited because of participation and reporting bias of healthy participants and by the extent to which participants may be unaware of the medical history of family members.

Any bias in model inputs may also bias the outputs and influence disparities. Modeling teams can decrease the risk of building biases into the models by carefully considering the implications of model assumptions and structure, data input choices, and the resultant solutions suggested by model findings. For example, because the distribution of age at diagnosis for some cancer types skew younger for Black as compared with White patients with cancer, models may need to consider evaluating the potential for higher tumor initiation rates, shorter sojourn times, and/or faster tumor growth rates in Black persons. These natural history parameters should not be interpreted as reflecting that Black people are inherently biologically different from White people but that these parameters reflect the upstream (inherited genetics) and downstream consequences of factors that exert physiological effects on persons subjected to racism, social isolation, environmental exposures, and violence as well as modifiable behavioral factors (12). The lung cancer article in this Monograph demonstrates an example of how the impact of different natural history components (eg, histology, stage, survival) on racial disparities, shaped by race-specific parameter input data (eg, smoking patterns), can be quantified (24,25). Furthermore, when available, model inputs may have greater uncertainty for population groups with smaller sample sizes in source databases leading to correspondingly imprecise and inaccurate outputs. Indeed, the relative lack of data concerning natural history parameters for different populations is a barrier to building models that are well calibrated within specific population groups. Biases can be present in all types of data sources including those that appear objectively measured, such as cancer recurrence or tumor marker presence, as well as data elements that can be more subjective, such as quality of life, tobacco, or substance abuse—behaviors that rely on self-report (Table 1, patient-reported outcomes, survival). To decrease the risk of building models that incorporate biased assumptions or structures that lead to erroneous conclusions, CISNET models are increasingly adopting model components that reflect upstream factors and actionable levers rather than solely generating comparisons of racial disparities (Table 1) (12). Here, we seek to address data at points along the

Table 1. Intervention targets and data elements for addressing health equity using modeling of care delivery along the cancer control continuum

Place in cancer control continuum	Intervention targets associated with cancer outcomes[a] and amenable to disparities modeling research	Required frequency distributions or rates of data elements specific to the population of interest[b]
"Upstream" structural factors	Health, social, and economic policies; social and environmental factors	Income, education, health literacy, health insurance coverage, employment, medical debt, residential segregation and mortgage lending practices, neighborhood factors (resources, violence), environmental quality (air, water), voting participation, local media and advertising exposure
Prevention and risk assessment	Individual cancer risk prediction, risk reduction behaviors and policies, access to genetic counseling	Risk factors (eg, family history of cancer, smoking status, environmental and occupational exposures), genetic test results if conducted, availability of genetic counseling
Early detection	Modality of screening test, availability and affordability of screening including new modalities, hours facilities are open	Test performance values including rates of false-positive results and biopsies after false-positives, test uptake and adherence, distributions of distance to screening facilities and facility characteristics such as area segregation and insurance accepted
Diagnosis	Local and regional health-care capacity including transfer of care between primary and specialty clinicians, health-care facility availability, screening failures (eg, interval cancers and advanced-cancer diagnoses despite recommended screening)	Follow-up rates, completeness of workup, time to follow-up after a positive screening test, work leave policies, time to treatment initiation, distance to facilities, facility characteristics including clinic workflow, stage at diagnosis, subtype of cancer
Treatment	Availability and quality of health care, challenges in transition of care between primary and oncology care, insurance coverage, out-of-pocket costs of care, insurance network restrictions, preauthorization requirements, availability of tumor biomarker testing	Facility characteristics including clinic workflow, availability of patient navigation, insurance type, costs of care, medical debt, treatment effectiveness, completion of guideline-concordant care, treatment type and quality
Patient-reported outcomes	Treatment shared decision making, symptom management, care coordination	Quality of life (utilities), satisfaction with care, symptoms (eg, pain, sleep quality, fatigue), documentation of shared decision making, type and timing of physician appointments, community resources, social determinants of health (eg, food insecurity), social capital and support, resilience, availability of paid sick days, patient navigation
Survival	Behavioral risk factor modification, surveillance testing, availability of maintenance therapy, survivorship care plans	Risk factors assessed pre- and postdiagnosis, patterns of surveillance screening tests and cancer care for recurrence and new cancers, receipt of survivorship care per plan

[a] Outcomes produced by models include cancer-specific incidence, survival, and mortality; life-years and quality-adjusted life-years; stage distribution of cancers diagnosed; false-positive screening tests; overdiagnosed cases; and health-care costs.

[b] Simulation models use group-level summary data as parameter inputs and calibration targets. Summary data include frequency distributions (eg, percent of persons in each category of a factor) and other statistics such as means or medians, rate ratios, relative risks, hazard ratios, and 95% confidence intervals.

cancer control continuum and how dataset selection may influence model findings.

Data needs for model inputs and calibration targets

A first challenge for collecting model inputs and calibration targets is the lack of data for different populations (eg, lack of detailed information on race, ethnicity, sexual orientation, income, education, marital status, disability, and other important sociodemographic characteristics) in surveillance, administrative, and other databases. If information is indeed available, representation and generalizability, harmonization, and completeness and accuracy contribute to whether data are useful and of sufficiently high quality for population model inputs or calibration targets to investigate strategies for improving health equity (see Table 2 for details).

Representation and generalizability

Nationally representative data are generally sought for model inputs, particularly for groups not well represented in commonly used sources. The imperfect gold standard for national representation is set by census estimates of at-risk population sizes and death certificates for cancer mortality counts (Table 2), although even this source is subject to undercounting of many demographic groups, including those most at risk of systemic discrimination and racism (26). The application of these estimates, however, is often within health-related data settings that are limited to health systems serving populations covered by commercial health insurance, Medicare, Medicaid, or veterans (eg, Table 2; medical claims and electronic health records [EHRs] or cancer registries for geographic areas). For example, CISNET investigators have collaborated with members of the NCI-supported BCSC and Population-based Research to Optimize the Screening Process (PROSPR) consortia to examine questions related to disparities in the delivery of cancer-related health care (27). The data collected by the PROSPR research centers reflect the variation of US delivery system organizations. However, some health-care–derived data sets inherently overrepresent individuals with health insurance and greater health-care access. Because access to health care is on the causal pathway between

Table 2. Strengths and limitations of data resources for modeling cancer equity

Data resource	Data elements relevant to modeling	Strengths	Limitations[a]
Customized data summaries from research studies like the Multi-Ethnic Cohort	Risk factor distributions according to key demographic factors such as age, sex, race, and ethnicity	Self-reported data for elements not routinely captured in medical records; often enriched for populations of interest	Missing data selectively more likely in certain groups, social biases and stigma affect certain groups differently when self-reporting, healthy cohort bias
Medicare	Patterns of medical care among persons aged 65 years and older and those with disabilities according to key demographic factors	Large, nationally representative sample (98% of those aged 65 years and older), potential for linkage to other detailed datasets (census, state cancer registries, National Death Index, provider information)	Excludes health maintenance organization patients and patients aged younger than 65 years; data more likely to be incomplete for individuals from disenfranchised populations; lack of data on underdiagnosed conditions, risk factors, and exposures
Medicaid (administrative claims)	Patterns of medical care among eligible low-income adults, children, pregnant women, elderly adults, and people with disabilities	Large, nationally representative sample (approximately 20% of the US population); data on those aged younger than 65 years	As a joint federal–state-funded program, eligibility, coverage, and scope varies across states and time necessitating national and state level analyses; substantial data lags (≥4 years)
Medical claims databases (eg, MarketScan)	Patterns of medical care according to key demographic factors	Details of full course of care including dosing and rounds of therapy not available in cancer registries	Limited to persons with certain types of health insurance, procedure data lack reason for service (eg, screening vs diagnostic follow-up), lack of data on risk factors and exposures
Public health surveillance systems (eg, American Community Survey, BRFSS, Census, MEPS, NHANES, NHIS, PATH)	Population size of United States by age, calendar year, and birth cohort, risk factor prevalence over time and by demographic characteristics	Nationally representative	Missing data selectively more likely in certain groups, social biases and stigma affect certain groups differently when self-reporting
State and national cancer registries supported by CDC and NCI (eg, SEER)	Cancer incidence, survival, and mortality; stage distribution at diagnosis; tumor factors such as grade and subtype	Near complete for geographic catchment areas, databases linked to SEER (Medicare, Medicaid, health outcomes, consumer assessment of providers)	Limited information on race and ethnicity before 1990s, treatment information often limited to planned first course, limited data on individual exposures and risk factors
Electronic medical records (eg, PROSPR, BCSC)	Patterns of medical care according to key demographic factors, individual-level risk prediction score values	Often enhanced through linkage with surveys or geospatial data, detailed data across the cancer care continuum including test results, diversity of types of health-care systems	Missing or incomplete data for patients with barriers to health-care access, limited data on risk factors and exposures
Clinical cooperatives (eg, NCDB, NCCN)	Patterns of cancer treatment care according to key demographic and disease factors	Can be very large and detailed for certain aspects of care	Limited inclusion of diverse populations
Death certificates (eg, CDC Wonder, Human Mortality Database)	Underlying cause of death, may include contributing causes	Near complete for the United States	Subject to well-described errors and bias for cause of death and race and ethnicity
Published literature including meta-analyses	Treatment efficacy according to disease factors	Treatment benefit under ideal clinical trial conditions	Restrictive study inclusion and exclusion criteria limit the diversity of the participating population

[a] Common limitation to all data sources: no information collected on many factors (country of origin, sex and gender minority status, sexual orientation, Veteran status, etc.); small numbers of race and ethnic groups other than non-Hispanic White; race and ethnicity historically not self-reported. BCSC = Breast Cancer Surveillance Consortium; BRFSS = Behavioral Risk Factor Surveillance System; CDC = Center for Disease Control and Prevention; MEPS = Medical Expenditure Panel Survey; NCCN = National Comprehensive Cancer Network; NCDB = National Cancer Database; NCI = National Cancer Institute; NHANES = National Health and Nutrition Examination Survey; NHIS = National Health Interview Survey; PATH = Population Assessment of Tobacco and Health Study; PROSPR = Population-based Research to Optimize the Screening Process; SEER = Surveillance, Epidemiology, and End Results.

racism and outcomes, data are more likely to be incomplete for individuals who are excluded from obtaining health care because of the effects of racism and discrimination. Alternatively, some health-care datasets, such as those from Medicaid or safety net systems, may reflect unique populations with less health-care access and local differences in eligibility. To improve representation of individuals often excluded from health studies and EHRs, data are especially needed that better represent the demographic characteristics, health behaviors, and clinical factors among persons with no or limited interaction with health-care services (Table 3).

For a given model input, an ideal data source would include relatively complete ascertainment of the relevant outcome measure for the entire population, not only individuals with unrepresentative health-care access. For instance, in comparison with EHR data, population-based registry data provide a better source for ascertaining natural history of some diseases because they include more information for the entire population within their geographic catchment area and not only those accessing health care. Yet, EHRs and other nonrepresentative data can also provide important information to inform models, which later can be extrapolated to the whole US population via model calibration and validation. In addition, supplementing registries, EHRs, and other data sources with data reflecting measures of systemic racism or its secondary effects could provide new sources of data for models to obtain more nuanced estimates of impacts on cancer types, stage, and response to therapy.

Harmonization

Data collected following a standardized protocol over a long period of time are helpful for examining trends and projecting outcomes resulting from interventions that may reverse trends that reflect disparities. Surveillance studies supported by the National Center for Health Statistics within the Centers for Disease Control and Prevention have served this essential purpose for decades (Table 2). The conversion of medical records from paper to electronic formats greatly expanded capability for health research and cancer modeling. However, health systems and clinics across the country have converted to EHRs at different times with varying capacity to support analysis. Furthermore, health systems have only recently started collecting self-reported race and ethnicity, whereas other characteristics, such as sexual orientation, gender identity, income, education, other social determinants of health, and individual experiences of discrimination, are not routinely ascertained [see Jayasekera et al. (28) in this Monograph]. Conversely, residential addresses are commonly collected and can be used to estimate area-level metrics of the effects of structural racism such as neighborhood disinvestment and disadvantage. Data sources with detailed individual level information on experiences and intermediate effects related to racism as well as health outcomes would allow population models to connect the dots from potential upstream drivers to downstream impacts (28,29). Models can then be reprogrammed to better include modifiable targets influenced by racism, including area-level metrics, social determinants of health, health insurance policy, environmental exposures, education quality, income, and debt (Table 3).

Because PROSPR has invested in harmonizing key EHR data elements for research purposes by identifying which elements are conceptually equivalent and can be pooled across medical records from different health systems, PROSPR data are conducive to use in population models as parameter inputs (30). For example, CISNET and PROSPR have collaborated to examine the impact of differences in time after an abnormal screening test to diagnostic evaluation on mortality from breast, cervical, and colorectal cancer (31). This study relied on standardized definitions and data capture for abnormal test results and time to diagnostic evaluation. The CISNET models including these harmonized measures can then be employed to estimate the potential long-term health effects of differences in time to care, incorporating any differences that may be experienced disproportionately by persons not historically represented in research.

Completeness and accuracy

Missing and inaccurate data are common and can be associated with cancer outcomes, including risk factors such as smoking status as well as tumor stage and subtype at diagnosis (32). If missing or misclassified values are more common among groups experiencing health disparities related to barriers to health care, then modeling results will be more prone to error or bias for these groups. Strategies to address these potential limitations can include sensitivity analyses to explore a range of input values, yet a single best value (or range of values) needs to be selected for each base-case modeling analysis. Data from epidemiologic cohort studies can provide self-reported information that is not routinely included in medical records or other sources (Table 2). For example, the CISNET lung modelers are collaborating with the Multiethnic Cohort to obtain input parameter data to connect individual smoking histories to lung cancer risk among Black, non-Hispanic White, and Hispanic adults as well as adults overall (24); this effort builds on early work based on the Nurses' Health Study and Health Professionals' Follow-up Study, which are predominately composed of non-Hispanic White adults (33). Reviews of published studies have suggested that the self-report of a cancer screening test is often more accurate when compared with medical records among persons who have received the test and less accurate among persons who have never received the test (34-36). Errors in relying on self-reported smoking and screening information may vary across populations because of cultural differences in a person's likelihood to acquiesce (respond yes when uncertain) and answer according to social desirability (overreport events that are socially favored) (34). Moreover, risk factors that are disproportionately relevant to underrecognized populations might have lower priority for accuracy in data collection efforts than those affecting most of the population. An example is menthol cigarette and cigar use, which is more common among Black individuals but less common in national and other surveys than other tobacco products that are reported less frequently in the Black population, such as e-cigarette use (37). Further, geocoding of residential addresses with linkage to area-level data can complement or substitute for self-reported data but also introduces other sources of error; in some situations, geocoding requires substantial investment and has limited accuracy, for example, for persons living in rural areas, temporary housing, mobile homes, and homeless shelters, and who use post offices boxes. The intersectionality of race, ethnicity, and access to care make patterns of exposure difficult to disentangle from the drivers of reporting accuracy. Thus, modeling research needs to consider the potential impact of multiple sources of bias to avoid overstating screening test use and exposure to healthy and unhealthy behaviors based on self-reports and to encourage the increased acquisition of accurate data on factors that may be overrepresented in populations historically and intentionally excluded from research.

Other opportunities remain to improve the systematic collection of model data inputs across all populations who are included in clinical databases (Table 3). Because EHRs were designed for billing and clinical purposes rather than for research, critical data are often missing or described in text notes rather than discrete fields (38). For example, smoking history (ie, pack-years and cessation data) is needed for identification of patients eligible for lung cancer screening and is often missing from or incomplete in the EHR (39,40). Informatics tools like character recognition and artificial intelligence are rapidly expanding the opportunities for clinical research based on unstructured EHR data, and novel methods for EHR-based phenotyping have been developed to improve the quality of characterization of populations in the presence of the many data challenges noted above. Policies and processes for protecting confidentiality and patient privacy will need to keep pace with these technology innovations (41).

Table 3. Priorities for improving data inputs used by population models of cancer equity

Limitation of current data	Priorities for future data resources
Omission of data from people with limited use of health care and participation in research studies	Cancer risk factor information, patterns of care, and health outcomes among uninsured and underinsured groups and other disenfranchised populations living in geographically diverse areas
Lack of data on exposures relevant for underrecognized groups	Identify relevant exposures underrepresented in national surveys and other data sources; expand questionnaires of national surveys to capture relevant risk factors
Coarse or broad race and ethnicity categories that conceal important variation in health outcomes	Disaggregate race and ethnic groups, especially for Alaska Native, Asian American, Hispanic, Native American, and Pacific Islander persons; gather necessary information to characterize persons at the intersection of multiple identities including those who identify as multiracial
Poor data quality or unavailable data identifying disenfranchised populations in clinical databases	Include standard discrete data with robust data confidentiality protections for race and ethnicity, sex, and gender beyond male and female, disability, sexual orientation, immigration, incarceration, and language preference
Unknown eligibility for screening tests and method of cancer detection; unclear whether tests are for screening or diagnostic follow-up	Include risk factor data to determine eligibility for cancer screening tests (eg, pack-years of smoking), method of detection (symptoms, screening), and purpose of tests (screening or diagnostic follow-up) in cancer registries and medical records
Self-reported data that are susceptible to misclassification and sampling bias; lack of individual-level data on social determinants of health in clinical datasets	Facilitate geocode linkages between medical records and claims with area-based measures of social determinants of health; facilitate linkage between surveys and medical records or claims for individual-level measures of social determinants of health, for example, National Health Interview Survey, National Health and Nutrition Examination Survey, and Medical Expenditure Panel Survey linked with Medicare; strengthen health information technology policies and procedures that allow data linkages while preserving patient confidentiality
Absence of data on measures of systemic racism	Increase availability of data on factors that reflect the effects of systemic racism including income, education, health literacy, employment, health insurance coverage, medical debt, residential segregation and mortgage lending practices, neighborhood factors (resources, violence), environmental quality (air, water), voting participation, local media and advertising exposure, and individual experiences of discrimination

Because the EHRs will likely always lack key data elements across the cancer continuum, enhanced EHR-based research data resources will continue to serve as a critical source of data inputs for modeling cancer disparities and health equity strategies. Enhancements include linkages to surveys and other databases using patient identifiers and to area-based measures using latitude–longitude geocodes (42,43).

Sample size

To adequately characterize cancer screening and treatment utilization patterns, as well as natural history of disease and history of exposures, an adequate sample size is essential for precise estimates, particularly for smaller racial and ethnic populations with diverse socioeconomic indicators. Underrepresentation of minoritized groups in clinical trials and other research studies resulting in imprecise parameter estimation and limited ability to obtain high-quality model inputs for these groups has been described extensively in the literature on algorithmic fairness, defined as "the study of definitions and methods related to the justice of models" (44). Combining data across multiple studies, health-care systems, or claims databases is often necessary to obtain adequate representation of individuals subject to potential health-care disadvantage. Collaborative data enterprises, such as those led by the National Comprehensive Cancer Network and the Commission on Cancer, have proved to be valuable as a source of data inputs for treatment patterns and clinical characteristics of Alaska Native, American Indian, Black, and Hispanic cancer patients (Table 2). As CISNET teams increase their capability to model race groups other than Black and White, ethnicity, and potentially intervenable targets that could modify the effects of systemic racism, corresponding new data inputs are needed. Ideally, new data sources would allow the disaggregation of race and ethnicity categories that combine heterogeneous groups, including Alaska Native, Asian American, Hispanic, Native American, and Pacific Islander persons (Table 3) (45,46). As reflected in other articles in this Monograph (24,47-51), CISNET modelers have increased efforts to examine cancer control strategies in Black persons. Additional efforts are needed to appropriately represent other underrecognized populations who experience racism and structural barriers to access health care, including groups at the intersection of multiple identities (52,53). Data harmonized through independent or collaborative efforts have also served as an important alternative to dependency on research reports and meta-analyses that may lack treatment efficacy estimates, for example, for racial and ethnic populations (Table 2) (54). Although smaller sample sizes may increase variability around model input values, concerns about data quality should not be used as an excuse to avoid health equity modeling research. Instead, investment in novel data resources should be pursued along with targeted sensitivity analyses to explore the impact of input data variation.

Future directions

Simulation modeling of the population cancer burden can be a powerful tool for identifying approaches to improve health equity and reduce cancer disparities, although we caution that existing data sources and modeling approaches can incorporate and perpetuate the effects of systemic racism and other upstream causes of disparities. Modeling teams including informaticists who advise and participate in those teams can take several steps to improve health equity modeling and limit the impact of underlying biases in data, including identifying data sources representative of the target population for modeling; developing new data sources that incorporate measures of systemic racism and its effects; decreasing missing data and misclassified data from populations historically and intentionally excluded from research; understanding the limits of existing data inputs; and using sensitivity analyses to estimate the effects of these limits on outcomes and conclusions. The NCI has heavily invested in data and consortium resources, including the BCSC and PROSPR, which draw on key data elements supplemented with additional data linkages. Building such large population models of cancer using the best combination of data inputs is an approach for identifying strategies to reduce the risk of excessive harm to population groups who have already suffered because of underrepresentation in medical research and not fully received its benefits. New multidisciplinary research teams are encouraged to develop population models and explore new and emerging data resources that may close persistent knowledge gaps resulting from the under- or misrepresentation of some people in existing models and data (55). Combined, these efforts can improve our ability to develop and use population models to evaluate health disparities, identify leverage points to modify contributing socioeconomic and health policies, and ultimately improve health equity.

Data availability

No new data were generated or analyzed in support of this research.

Author contributions

Amy Trentham-Dietz, PhD (Conceptualization; Funding acquisition; Writing—original draft; Writing—review & editing), Douglas A. Corley, MD, PhD (Conceptualization; Writing—review & editing), Natalie J. Del Vecchio, PhD (Conceptualization; Writing—review & editing), Robert T. Greenlee, PhD, MPH (Conceptualization; Writing—review & editing), Jennifer S. Haas, MD, MSc (Conceptualization; Writing—original draft; Writing—review & editing), Rebecca A. Hubbard, PhD (Conceptualization; Writing—original draft; Writing—review & editing), Amy E. Hughes, PhD (Conceptualization; Writing—review & editing), Jane J. Kim, PhD (Conceptualization; Writing—original draft; Writing—review & editing), Sarah Kobrin, PhD, MPH (Conceptualization; Writing—original draft; Writing—review & editing), Christopher I. Li, MD, PhD (Conceptualization; Writing—review & editing), Rafael Meza, PhD (Conceptualization; Writing—review & editing), Christine M. Neslund-Dudas, PhD (Conceptualization; Writing—review & editing), and Jasmin A. Tiro, PhD, MPH (Conceptualization; Writing—original draft; Writing—review & editing).

Funding

This work was supported by the National Institutes of Health under National Cancer Institute Grant U01CA253911. This research was also supported in part by the National Institutes of Health under National Cancer Institute Grants P30CA014520, UM1CA221940, P01CA154292, UM1CA222035, U24CA221936, U01CA152637, and UM1CA221939, and the American Cancer Society (CRP-22-080-01-CTPS).

The funders did not influence this publication or review previous drafts.

The views expressed here are those of the authors only and do not necessarily reflect the views of the National Cancer Institute or National Institutes of Health.

Monograph sponsorship

This article appears as part of the monograph “Reducing Disparities to Achieve Cancer Health Equity: Using Simulation Modeling to Inform Policy and Practice Change,” sponsored by the National Cancer Institute, National Institutes of Health ([Comparative Modeling of Precision Breast Cancer Control Across the Translational Continuum; 3 U01 CA253911-03S2]).

Conflicts of interest

None.

Acknowledgments

The authors thank Drs. Jeanne Mandelblatt and Aruna Kamineni for editorial suggestions.

References

1. Holford TR, Meza R, Warner KE, et al. Tobacco control and the reduction in smoking-related premature deaths in the United States, 1964-2012. *JAMA*. 2014;311(2):164-171.
2. Jeon J, Holford TR, Levy DT, et al. Smoking and lung cancer mortality in the United States from 2015 to 2065: a comparative modeling approach. *Ann Intern Med*. 2018;169(10):684-693.
3. Burger EA, Smith MA, Killen J, et al. Projected time to elimination of cervical cancer in the USA: a comparative modelling study. *Lancet Public Health*. 2020;5(4):e213-e222.
4. Berry DA, Cronin KA, Plevritis SK, et al.; Cancer Intervention and Surveillance Modeling Network (CISNET) Collaborators. Effect of screening and adjuvant therapy on mortality from breast cancer. *N Engl J Med*. 2005;353(17):1784-1792.
5. Plevritis SK, Munoz D, Kurian AW, et al. Association of screening and treatment with breast cancer mortality by molecular subtype in US women, 2000-2012. *JAMA*. 2018;319(2):154-164.
6. Etzioni R, Gulati R, Tsodikov A, et al. The prostate cancer conundrum revisited: treatment changes and prostate cancer mortality declines. *Cancer*. 2012;118(23):5955-5963.
7. Mandelblatt JS, Stout NK, Schechter CB, et al. Collaborative modeling of the benefits and harms associated with different U.S. breast cancer screening strategies. *Ann Intern Med*. 2016; 164(4):215-225.
8. Knudsen AB, Rutter CM, Peterse EFP, et al. Colorectal cancer screening: an updated modeling study for the US Preventive Services Task Force. *JAMA*. 2021;325(19):1998-2011.
9. Meza R, Jeon J, Toumazis I, et al. Evaluation of the benefits and harms of lung cancer screening with low-dose computed tomography: modeling study for the US Preventive Services Task Force. *JAMA*. 2021;325(10):988-997.
10. Kim JJ, Burger EA, Regan C, et al. Screening for cervical cancer in primary care: a decision analysis for the US Preventive Services Task Force. *JAMA*. 2018;320(7):706-714.
11. Lowry KP, Geuzinge HA, Stout NK, et al.; Breast Working Group of the Cancer Intervention and Surveillance Modeling Network (CISNET), in collaboration with the Breast Cancer Surveillance Consortium (BCSC), and the Cancer Risk Estimates Related to Susceptibility (CARRIERS) Consortium. Breast cancer screening

strategies for women with ATM, CHEK2, and PALB2 pathogenic variants: a comparative modeling analysis. *JAMA Oncol*. 2022;8(4):587-596.

12. Chapman C, Jayasekera J, Dash C, et al. A health equity framework to support the next generation of cancer population simulation models. *J Natl Cancer Inst Monogr*. 2023.
13. Kaur D, Ulloa-Perez E, Gulati R, et al. Racial disparities in prostate cancer survival in a screened population: Reality versus artifact. *Cancer*. 2018;124(8):1752-1759.
14. Chapman CH, Schechter CB, Cadham CJ, et al. Identifying equitable screening mammography strategies for black women in the United States using simulation modeling. *Ann Intern Med*. 2021;174(12):1637-1646.
15. Han SS, Chow E, Ten Haaf K, et al. Disparities of national lung cancer screening guidelines in the US population. *J Natl Cancer Inst*. 2020;112(11):1136-1142.
16. Rutter CM, May FP, Coronado GD, et al. Racism is a modifiable risk factor: relationships among race, ethnicity, and colorectal cancer outcomes. *Gastroenterology*. 2022;162(4):1053-1055.
17. Campos NG, Scarinci IC, Tucker L, et al. Cost-effectiveness of offering cervical cancer screening with HPV self-sampling among African-American women in the Mississippi delta. *Cancer Epidemiol Biomarkers Prev*. 2021;30(6):1114-1121.
18. Beaber EF, Kim JJ, Schapira MM, et al.; on behalf of the Population-based Research Optimizing Screening through Personalized Regimens consortium. Unifying screening processes within the PROSPR consortium: A conceptual model for breast, cervical, and colorectal cancer screening. *J Natl Cancer Inst*. 2015;107(6):djv120.
19. Beaber EF, Kamineni A, Burnett-Hartman AN, et al. Evaluating and improving cancer screening process quality in a multilevel context: the PROSPR II consortium design and research agenda. *Cancer Epidemiol Biomarkers Prev*. 2022;31(8):1521-1531.
20. Trentham-Dietz A, Alagoz O, Chapman C, et al.; Breast Working Group of the Cancer Intervention and Surveillance Modeling Network (CISNET). Reflecting on 20 years of breast cancer modeling in CISNET: recommendations for future cancer systems modeling efforts. *PLoS Comput Biol*. 2021;17(6):e1009020.
21. Bowleg L. The problem with the phrase women and minorities: intersectionality-an important theoretical framework for public health. *Am J Public Health*. 2012;102(7):1267-1273.
22. Shiyanbola OO, Arao RF, Miglioretti DL, et al. Emerging trends in family history of breast cancer and associated risk. *Cancer Epidemiol Biomarkers Prev*. 2017;26(12):1753-1760.
23. Murff HJ, Peterson NB, Greevy R, et al. Impact of patient age on family cancer history. *Genet Med*. 2006;8(7):438-442.
24. Skolnick S, Cao P, Jeon J, et al. Contribution of smoking patterns, disease natural history, and survival on lung cancer disparities in non-Hispanic Black individuals: a modeling study. *J Natl Cancer Inst Monogr*. 2023.
25. Meza R, Cao P, Jeon J, et al. Patterns of birth cohort–specific smoking histories by race and ethnicity in the U.S. *Am J Prev Med*. 2023;64(4 suppl 1):S11-S21.
26. Jensen E, Kennel T. *Who Was Undercounted, Overcounted in the 2020 Census?*, March 10, 2022. https://www.census.gov/library/stories/2022/03/who-was-undercounted-overcounted-in-2020-census.html. Accessed April 21, 2023.
27. Kim JJ, Tosteson AN, Zauber AG, et al. Cancer Models and Real-world Data: Better Together: Table 1. *J Natl Cancer Inst*. 2016;108(2):djv316.doi:10.1093/jnci/djv316
28. Jayasekera J, Fernandes JR, Woo JMP, et al. Opportunities, challenges, and future directions for modeling the effects of structural racism on cancer mortality in the U.S.: a scoping review. *J Natl Cancer Inst Monogr*. 2023.
29. Ray R, Lantz PM, Williams D. Upstream policy changes to improve population health and health equity: a priority agenda. *Milbank Q*. 2023;101(S1):20-35.
30. Healthcare Delivery Research Program, Division of Cancer Control & Population Sciences, National Cancer Institute. *PROSPR DataShare*. April 10, 2023. https://healthcaredelivery.cancer.gov/prospr/datashare/. Accessed April 21, 2023.
31. Rutter CM, Kim JJ, Meester RGS, et al. Effect of time to diagnostic testing for breast, cervical, and colorectal cancer screening abnormalities on screening efficacy: a modeling study. *Cancer Epidemiol Biomarkers Prev*. 2018;27(2):158-164.
32. Yang DX, Khera R, Miccio JA, et al. Prevalence of missing data in the national cancer database and association with overall survival. *JAMA Netw Open*. 2021;4(3):e211793.
33. Meza R, Hazelton WD, Colditz GA, et al. Analysis of lung cancer incidence in the Nurses' Health and the Health Professionals' Follow-Up Studies using a multistage carcinogenesis model. *Cancer Causes Control*. 2008;19(3):317-328.
34. Anderson J, Bourne D, Peterson K, et al. *Evidence Brief: Accuracy of Self-Report for Cervical and Breast Cancer Screening*. VA ESP Project #09-199. Washington, DC: US Department of Veterans Affairs; 2019. https://www.ncbi.nlm.nih.gov/books/NBK539386/pdf/Bookshelf_NBK539386.pdf. Accessed April 21, 2023.
35. Howard M, Agarwal G, Lytwyn A. Accuracy of self-reports of Pap and mammography screening compared to medical record: a meta-analysis. *Cancer Causes Control*. 2009;20(1):1-13.
36. Rauscher GH, Johnson TP, Cho YI, et al. Accuracy of self-reported cancer-screening histories: a meta-analysis. *Cancer Epidemiol Biomarkers Prev*. 2008;17(4):748-757.
37. Zavala-Arciniega L, Meza R, Hirschtick JL, et al. Disparities in cigarette, e-cigarette, cigar, and smokeless tobacco use at the intersection of multiple social identities in the US adult population. Results from the tobacco use supplement to the current population survey 2018-2019 survey. *Nicotine Tob Res*. 2023;25(5):908-917.
38. Taksler GB, Dalton JE, Perzynski AT, et al. Opportunities, pitfalls, and alternatives in adapting electronic health records for health services research. *Med Decis Making*. 2021;41(2):133-142.
39. Modin HE, Fathi JT, Gilbert CR, et al. Pack-year cigarette smoking history for determination of lung cancer screening eligibility. Comparison of the electronic medical record versus a shared decision-making conversation. *Ann Am Thorac Soc*. 2017;14(8):1320-1325.
40. Ritzwoller DP, Meza R, Carroll NM, et al. Evaluation of population-level changes associated with the 2021 US Preventive Services Task Force lung cancer screening recommendations in community-based health care systems. *JAMA Netw Open*. 2021;4(10):e2128176.
41. Elmore LW, Greer SF, Daniels EC, et al. Blueprint for cancer research: critical gaps and opportunities. *CA Cancer J Clin*. 2021;71(2):107-139.
42. Shih YT, Sabik LM, Stout NK, et al. Health economics research in cancer screening: research opportunities, challenges, and future directions. *J Natl Cancer Inst Monogr*. 2022;2022(59):42-50.
43. Doria-Rose VP, Breen N, Brown ML, et al. A history of health economics and healthcare delivery research at the National Cancer Institute. *J Natl Cancer Inst Monogr*. 2022;2022(59):21-27.

44. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. *Annu Rev Biomed Data Sci*. 2021;4:123-144.
45. Kanaya AM, Hsing AW, Panapasa SV, et al. Knowledge gaps, challenges, and opportunities in health and prevention research for Asian Americans, native Hawaiians, and Pacific Islanders: a report From the 2021 National Institutes of Health Workshop. *Ann Intern Med*. 2022;175(4):574-589.
46. Cancer Disparities Progress Report 2022: Achieving the bold vision of health equity for racial and ethnic minorities and other underserved populations. 2022. http://www.CancerDisparitiesProgressReport.org/. Accessed April 21, 2023.
47. Chapman CH, Schechter CB, Huang H, et al. Racial disparities in US breast cancer mortality. *J Natl Cancer Inst Monogr*. 2023.
48. Gulati R, Nyame YA, Lange JM, et al. A model-based decomposition of racial disparities in prostate cancer incidence and mortality. *J Natl Cancer Inst Monogr*. 2023.
49. Rutter CM, Nascimento de Lima P, May FP, et al. Understanding racial disparities in colorectal cancer outcomes. *J Natl Cancer Inst Monogr*. 2023.
50. Sereda Y, Alarid-Escudero F, Bickell NA, et al. Approaches to developing de novo cancer population models to examine racial disparities in bladder, gastric, and endometrial cncer and multiple myeloma mortality: The CISNET incubator program. *J Natl Cancer Inst Monogr*. 2023.
51. Spencer JC, Burger EA, Campos NG, et al. Adapting a model of cervical carcinogenesis among self-identified Black women to evaluate racial disparities in the United States. *J Natl Cancer Inst Monogr*. 2023.
52. Kelly-Brown J, Palmer Kelly E, Obeng-Gyasi S, et al. Intersectionality in cancer care: A systematic review of current research and future directions. *Psychooncology*. 2022;31(5):705-716.
53. Malone J, Snguon S, Dean LT, et al. Breast cancer screening and care among black sexual minority women: a scoping review of the literature from 1990 to 2017. *J Womens Health (Larchmt)*. 2019;28(12):1650-1660.
54. Tam J, Levy DT, Feuer EJ, et al. Using the past to understand the future of U.S. and global smoking disparities: a birth cohort perspective. *Am J Prev Med*. 2023;64(4 suppl 1):S1-S10.
55. Meza R, Jeon J. Invited Commentary: Mechanistic and Biologically Based Models in Epidemiology-A Powerful Underutilized Tool. *Am J Epidemiol*. 2022;191(10):1776-1780.