Published in final edited form as:
Cancer Causes Control. 2022 June ; 33(6): 899–911. doi:10.1007/s10552-022-01571-0.

An informatics infrastructure to catalyze cancer control research and practice

Johnie Rose[1,2], **Weichuan Dong**[2,3], **Uriel Kim**[1,3], **Joseph Hnath**[1], **Abby Statler**[2,4], **Paola Saroufim**[5], **Sunah Song**[5], **Mustafa Ascha**[3,5], **Harry Menegay**[5], **Ye Tian**[5], **Mark Beno**[5], **Siran M. Koroukian**[2,3]

[1]Case Western Reserve University Center for Community Health Integration, 11000 Cedar Ave., Ste. 402, Cleveland, OH 44106-7136, USA

[2]Case Comprehensive Cancer Center, Cleveland, OH, USA

[3]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH, USA

[4]Taussig Cancer Institute, The Cleveland Clinic Foundation, Cleveland, OH, USA

[5]Cleveland Institute for Computational Biology, Case Western Reserve University/University Hospitals Cleveland Medical Center, Cleveland, OH, USA

Abstract

Purpose—A disconnect often exists between those with the expertise to manage and analyze complex, multi-source data sets, and the clinical, social services, advocacy, and public health professionals who can pose the most relevant questions and best apply the answers. We describe development and implementation of a cancer informatics infrastructure aimed at broadening the usability of community cancer data to inform cancer control research and practice; and we share lessons learned.

Methods—We built a multi-level database known as The Ohio Cancer Assessment and Surveillance Engine (OH-CASE) to link data from Ohio's cancer registry with community data from the U.S. Census and other sources. Space—and place-based characteristics were assigned to individuals according to residential address. Stakeholder input informed development of an interface for generating queries based on geographic, demographic, and disease inputs and for outputting results aggregated at the state, county, municipality, or zip code levels.

Results—OH-CASE contains data on 791,786 cancer cases diagnosed from 1/1/2006 to 12/31/2018 across 88 Ohio counties containing 1215 municipalities and 1197 zip codes. Stakeholder feedback from cancer center community outreach teams, advocacy organizations,

✉Johnie Rose, Johnie.rose@case.edu.

Author contributions JR, WD, UK, JH, UK, and SK contributed to project conception and design, as well as data acquisition. Database and interface development were performed by PS, SS, MA, HM, YT, and MB. JR drafted the manuscript, and all authors provided input. JR and SK obtained funding to support the work. All authors read and approved the final manuscript.

Code availability Code is available upon request.

Conflict of interest The authors report no conflicts of interest/competing interests.

Ethical approval This work was approved by the Institutional Review Boards of Case Western Reserve University and the Ohio Department of Health, and is governed by a data use agreement with the Ohio Department of Health for the use of cancer registry data.

public health, and researchers suggests a broad range of uses of such multi-level data resources accessible via a user interface.

Conclusion—OH-CASE represents a prototype of a transportable model for curating and synthesizing data to understand cancer burden across communities. Beyond supporting collaborative research, this infrastructure can serve the clinical, social services, public health, and advocacy communities by enabling targeting of outreach, funding, and interventions to narrow cancer disparities.

Keywords

Cancer control; Cancer prevention; Disparities; Database; Community-partnered research; Cancer registry

Introduction

Among the themes identified in the 2017 Unified Strategy Statement for Health Disparities Research from the American Association for Cancer Research (AACR), American Cancer Society (ACS), American Society of Clinical Oncology (ASCO), and U.S. National Cancer Institute (NCI) [1] was the need for understanding the interplay of factors at multiple levels —from molecular and clinical to socioeconomic and environmental—in driving differences in cancer outcomes across communities. The 2017 statement further identified the need for timely feedback of actionable data to key stakeholders as part of a cycle of information, action, and monitoring [1]. This feedback can empower community partners to better direct their interventions and advocacy. In addition, this process of feeding back data can foster a shared approach to posing research questions and interpreting results as part of community-partnered research [2, 3]. When deep understanding of methods and data limitations, as possessed by many in academia, meets the operational experience, cultural familiarity, and social capital of partners embedded in the community, powerful synergies can emerge. These synergies offer the potential to build academic understanding of cancer disparities while enhancing operational effectiveness of community partners who are fighting the cancer control and prevention battle on the ground.

The Public Health Informatics Institute (PHII) and Institute for Alternative Futures (IAF) recommends the creation of queryable data resources accessible by non-academic partners [4]. Federal [5, 6] and state [7] efforts to make county-level cancer burden and social determinants of health data available publicly are valuable public resources, but these portals may not provide the granularity needed for research, operational planning, outreach, and advocacy in community cancer control.

Here, we describe our approach to development of an evolving cancer informatics infrastructure in Ohio designed to support community-partnered cancer control research and to serve as a resource for stakeholders who need access to digestible, sub-county, multi-level cancer and community data to inform their activities. Importantly, this infrastructure was built using widely available data sources and primarily open source development tools. The prototype represents a model which is (a) expandable to include other individual-and community-level data sources and (b) transportable to other regions and states.

Methods

We built the Ohio Cancer Assessment and Surveillance Engine (OH-CASE), a Structured Query Language (SQL) relational database covering the state of Ohio. In addition, we built a point-and-click user interface which facilitates rapid searches of the data by those unfamiliar with SQL. The database is built using Microsoft SQL Server 2017, and the interface is built using RStudio (version 1.2.1335) and the R Shiny package (version 1.2.0). The choice of open source software for the interface is intended to foster sustainability and reproducibility. Figure 1 provides an overview of the OH-CASE architecture as described in the following sections.

In order to create a prototype that is reproducible in different locales, we used constituent data sets which are commonly available for most locations in the U.S. and many other countries. OH-CASE is housed in the *Secure Research Environment* (SRE) of Case Western Reserve University (CWRU). The SRE is a computing environment governed by a risk-based security program that includes implementation of controls meeting recommendations or requirements of regulatory and information security standards including the Health Insurance Portability and Accountability Act (HIPAA), the Federal Information Security Modernization Act (FISMA), and the International Organization for Standardization (ISO). Similar environments exist at other research institutions and academic medical centers.

Data sources

The following data sources are incorporated into the relational database model.

The Ohio Cancer Incidence Surveillance System (OCISS)—OCISS is Ohio’s National Program of Cancer Registries (NPCR) [8] supported and North American Association of Central Cancer Registries (NAACCR) [9] certified cancer registry. We obtained registry data for cases diagnosed from 2006 to 2018 (the most recent year for which data are available) under a data use agreement (DUA) with the Ohio Department of Health (ODH) and with approval of the Institutional Review Boards (IRBs) from ODH and Case Western Reserve University. OCISS contains fields for direct patient identifiers, demographics, insurance type, geocoded address at diagnosis, cancer site, stage, tumor size and histology, initial treatment modalities (including information on positive and negative lymph nodes excised), vital status (including cause of death, if applicable), occupation, and other domains. There are also fields for receptor status of breast cancers—the only cancer type for which reporting of tumor receptor status is required as part of public health cancer surveillance. The individual-level data in OCISS is linkable to any of the below community-level data sources based on geocoded patient address at diagnosis.

Five-year U.S. Census American Community Survey (ACS) data—The U.S. Census Bureau conducts an ongoing survey of Americans which is deployed to over 3 million households annually. This data is used to provide more timely estimates of key census parameters between decennial censuses [10]. We use published ACS data for five-year periods since estimates over shorter intervals would not provide figures for areas as small as census block groups [11]. These data provide contextual understanding of the demographic (age, race, ethnicity, sex, marital status) makeup, economic circumstances, housing environment, household composition, residential mobility/stability, educational attainment, and transportation resources of neighborhoods. In OH-CASE, we include five-year ACS data tables at the county, zip code tabulation area, census place (which we refer to as ‘municipality’), and census block group levels. A census block group is a subdivision of a census tract and contains anywhere from 600 to 3,000 people, most typically around 1,500; the census block group is the smallest unit for which sample-based estimates are available between decennial censuses [12]. In a city, this can represent an area of just a few city blocks.

Health Resources and Services Administration (HRSA) Health Professional Shortage Area (HPSA) data—This publicly available resource lists geographic divisions which are considered designated health professional shortage areas for primary care, mental health, and dental care based on provider to resident ratios. A HPSA score for each designated health professional shortage areas reflects the degree of shortage of the respective service type [13]. Linking to OCISS data based on geocoded address at diagnosis will allow identification of cancer patients living in HPSA-designated communities, and the degree of shortage based on HPSA score.

Food and Drug Administration (FDA) certified mammography facilities— Updated several times each year, this publicly available database contains the street addresses of all mammography facilities certified by the American College of Radiology or by states. Such certification is a requirement for providing mammography. [14] This data will allow quantification of mammogram density in a given geography or calculation of distance or travel time to mammography for members of a population of cancer patients.

Linkage

Geocoded patient address at the time of diagnosis is the basis for linking patient-level data from OCISS to the community-level data sources. Doing so requires creating a patient address shapefile using ArcGIS and performing a spatial join to match patient address to the appropriate shapefile corresponding to the geography of interest (zip code tabulation area, municipality, county, etc.). In the future, linkage of individual-level data from OCISS to new *individual-level* data sources will occur using a validated algorithm that utilizes combinations of individual identifiers as described by Koroukian et al. [15], except in the case of individual-level data from the Centers for Medicare and Medicaid Services (CMS). If linking to CMS data, we would need to provide individual-level identifiers to CMS after securing the necessary approvals, with CMS performing the linkage.

User interface development

We designed a user interface to allow stakeholder groups to easily search elements of the OH-CASE database to fulfill common use cases. The interface was designed with four types of users in mind: health system community outreach staff, advocacy organizations, public health agencies, and researchers. Meeting with representatives of each user type, we documented a series of proposed use cases that informed user interface design and prompted discussion of the types of search fields, results fields, and levels of data aggregation that would provide greatest utility to each group. During each meeting, we showed evolving prototype interfaces to the individual or group. We took notes and, at the end of each meeting, summarized the use cases and feedback provided by participants to ensure accuracy. We also received feedback through numerous one-on-one conversations with cancer researchers at our institution.

Results

OH-CASE contains 791,786 unique patient records for Ohio-ans diagnosed with cancer from 1/1/2006 to 12/31/2018. Of these, 0.25% had missing values for the key variables of address at diagnosis or diagnosis date. None was missing primary cancer site. The data spans all 88 Ohio counties, 1215 municipalities, 1197 zip codes, 2952 census tracts, and 9238 census block groups—an area containing 337 mammogram facilities. Table 1 provides summary statistics describing cancer cases represented in the most recent five years of OH-CASE data.

Stakeholder use cases

Local stakeholder input informed development of the use cases described in Table 2. To identify these use cases and obtain feedback on useful system features, we met with community outreach staff of two Cleveland, Ohio academic cancer hospitals (two meetings with five total staff), a community outreach specialist from a large public hospital (one meeting), the staff of two local cancer advocacy organizations (10 participants across two meetings), and a large local public health agency (five staff in one meeting). These meetings occurred over an approximately four-month period in mid-2019 as we designed and refined the OH-CASE Application interface.

Initial interface design—The initial version of the interface allows construction of queries and reporting of results based on OCISS and Census data. The following were integrated into the user interface design based on stakeholder input:

Geographic Aggregation Level-: Users are first asked to select a geographic level at which they would like query results to be aggregated: one group (statewide), by county, by zip code, or by municipality.

Filter Criteria-: Filter criteria from the categories below are applied to individual-level characteristics of OCISS patients. For any field besides date range, multiple options may be selected. Figure 2 shows the filter criteria fields of the OH-CASE interface.

- Start/End Dates—Users select a range of diagnosis dates. The interface does not offer the ability to specify exact beginning and ending dates. Specifying a very short date range in combination with other highly specific filter criteria could greatly increase the risk of re-identifying individuals. Instead, we chose to provide dropdown menus for selecting starting and ending *quarter*.
- Rural/urban continuum—Options are “In counties not lying in a metro area,” “In counties lying in a metro area of 250,000 to 1 million population,” and “In counties lying in a metro area of greater than 1 million population.”

- County—The user can restrict searches to certain counties. We included a preset option for the 15-county region corresponding to our National Cancer Institute (NCI) designated Cancer Center's northeast Ohio catchment area.
- Age at diagnosis—Categories corresponding to 14 years and younger, 15–29, 30–44, 45–54, 55–64, 65–74, 75–84, and 85 and older.
- Sex—Female, male, Transsexual/Not otherwise specified, and Other/biologically non-binary (This is the classification used in OCISS.)
- Race—American Indian or Alaska Native, Asian, Black or African American. Native Hawaiian or Other Pacific Islander, White or Caucasian, Other, and Unknown
- Ethnicity—Hispanic or Latino, Non-Hispanic/-Latino, and Ethnicity unknown
- Primary cancer site—We used standardized primary site definitions based on tumor location and histology as specified in the International Classification of Diseases for Oncology, Third Edition (ICD-O-3)/World Health Organization (WHO) 2008 definitions [16].
- SEER Summary Stage—Options are in situ (excluded by default except in the case of bladder cancer [16]), local, regional, distant, and unknown. OCISS requires reporting of Surveillance Epidemiology and End Results (SEER) Summary [17] but not American Joint Committee on Cancer (AJCC) cancer [18] stage. However, if one were to implement this model in a state where reporting of the more granular AJCC staging is mandatory, the staging filter criteria are easily changed accordingly.
- Breast Cancer Receptor Status—This field is activated if breast cancer is selected as the primary site. Options consist of combinations of estrogen receptor (ER), progesterone receptor (PR), and human epithelial growth factor receptor 2 (HER-2) status as follows:
  - ER/PR negative, HER2 negative (Triple Negative)
  - ER and/or PR positive, HER2 negative
  - ER/PR negative, HER2 positive
  - ER and/or PR positive, HER2 positive

A text box (not shown in Fig. 2) summarizes the search criteria applied, allowing the user to verify search parameters. In addition, a separate box at the bottom of the screen (also not shown) reproduces the actual SQL query statement executed based on the interface inputs. The latter is a useful mechanism by which system administrators can verify that correct queries are executed.

<u>Search Results—:</u> Search results are displayed with one row per selected geographic unit (i.e., one row for the entire state, one row per county, one row per zip code, or one row per municipality). The bottom three rows of each result set contain the following, respectively:

- Total—A summary row for results aggregated across all preceding rows. Counts are summed; rates and proportions are calculated from case and population totals.
- Ohio benchmark—The corresponding total, when available, from the compiled OCISS data for the entire state of Ohio
- National benchmark—The corresponding total, when available, from the SEER 99% sample data set [18]

Users may select the columns which will be displayed. The columns displayed by default are shown in Fig. 3 and fall into two broad categories as follows:

Case information

- Case counts—In addition to total case counts, crude case counts may be broken down (with percentages) by sex, age category, Black or African American (the predominant minority racial group in Ohio) and Non-African American race, and ethnicity.
- Age-adjusted Incidence rate—U.S. Census American Community Survey population totals based on the user-selected geographies and age range serve as denominators in the calculation of incidence rate per 100,000. For county and larger geographies, age-adjusted incidence and 95% confidence intervals are calculated using the direct method of age adjustment [19]; with the 2000 U.S. standard population as the reference, which is standard practice [20]. For municipalities and zip codes, age-adjusted incidence and 95% confidence intervals are calculated using the indirect method [19], referenced to Ohio-wide age-stratified incidence. Though the indirect method does not as readily allow comparison to rates calculated from different data sets, this method was chosen for smaller geographies because, unlike the direct method, it does not require age strata-specific case counts. The small numbers of cases in individual age strata within small and/or sparsely populated areas could lead to unstable estimates with wide confidence intervals if applying the direct method [19]. Age-adjusted incidence by race and ethnic categories may also be displayed.
- Stage—The proportion of cases diagnosed at a distant (metastatic) stage and with unknown stage are displayed by default for each geographic unit. A high proportion of distant stage disease may indicate poor access to screening or medical care generally. Previous work has shown that patients who are older, African American, have complex care needs, or are insured by Medicaid are more likely to have cancer of unknown stage [21, 22].
- Median time-to-treatment—Time to initial cancer treatment is operationalized as the date of tissue diagnosis subtracted from date of first treatment of any modality. For multiple cancer types, prolonged treatment delay may be a marker of low quality care or poor access [23].
- Proportion uninsured—This output is included as a marker of financial access to health services.

- Mortality rate—Age-adjusted mortality rate per 100,000 population is calculated, with direct age adjustment for county-level and larger geographies, and indirect adjustment for smaller geographies (as described for age adjustment of incidence). Median follow-up time is also included to provide context to the mortality figures.

Population Information

- Overall population counts as well as population counts broken down by sex, race, ethnicity, and age categories
- Median household income
- Proportion of the community living in high-Area Deprivation Index (ADI) census block groups—The ADI is a widely used index of social deprivation calculable from census variables. The index incorporates 17 separate factors covering domains of education, employment, income, housing (costs, crowding), and transportation access [24]. Deciles of ADI values are used where the highest possible ADI ranking is 10, representing the top ten percent of the most deprived block groups of all block groups in Ohio. As a scalable summary measure of deprivation which can be applied to any geography, we have included an algorithm which calculates the proportion of population in any geography living in census block groups within the ninth or tenth decile of ADI.

To prevent identification of individuals, and in accordance with our DUA with the Ohio Department of Health, any number or combination of numbers revealing a cell size of ten or fewer individuals is not displayed. Results are exportable as a comma-separated value (CSV) file.

Future versions of the OH-CASE Application will feature improved explanations of search fields and results column headings (via “mouseover” captions).

Real-world applications

The OH-CASE Database and Application have so far been applied in a number of real-world uses within our community and institutions:

- Quickly accessing population data for grant writing purposes, including understanding local burden and potential study population size as well as completing inclusion enrollment reporting for a population-based study
- Supporting description of catchment area cancer burden by leadership of our NCI-designated Comprehensive Cancer Center
- Creating publicly accessible data briefs regarding specific cancer types
- Enabling original research [25–27]

The Appendix provides two use case examples for how OH-CASE may provide insight into variations in cancer burden to inform research or focus the efforts of agencies seeking to mitigate cancer disparities.

Discussion

Here, we have described OH-CASE, a prototype multi-level cancer data infrastructure with a user interface designed to put sub-county-level cancer and community data into the hands of researchers and stakeholders who can apply the data in their local work. This prototype, designed with stakeholder input and built using data sources and tools which are widely available, can be implemented in diverse locales.

Such multi-level cancer data infrastructures offer the potential to achieve the widely espoused [1, 2, 4, 28, 29] goals of empowering local cancer control practitioners with data and enhancing the value and effectiveness of community-academic collaborations. Public and private community agencies leverage finite resources as they attempt to achieve maximal impact within their domains. Academic researchers strive to create generalizable knowledge about the extent and drivers of disparate cancer burdens across populations, while facing significant pressure to publish and maintain grant funding. While, ultimately, both groups share the same larger goals, the practical pressures on each create a misalignment. Community organizations have limited bandwidth to work with researchers, and there is a danger that these partners serve only as sources of data—a recipe for squandering social capital and damaging sustainability. We believe that accessible cancer and population data can serve as a nidus around which academic-community collaboration can coalesce, leading to better leveraging and application of complex data by community partners and to more contextually grounded research questions, better interpretation, and improved dissemination by researchers.

From conception to implementation, local action to reduce cancer disparities is most likely to be effective when it is data-driven but informed by the local wisdom residing in vulnerable communities [1, 2, 29]. For this reason, a number of academic-community partnerships have applied community-based participatory research (CBPR) principles to prioritizing community needs, identifying and implementing appropriate evidence-based interventions, and disseminating best practices [3, 30–33]. Academicians bring expertise in medicine, epidemiology, anthropology, and other fields to the table, providing varying degrees of structure, facilitation, and methodologic expertise. Community partners bring real-world expertise to the table based on their own lived experiences and those of the populations they serve. Their insights can shed light on the programmatic, practical, and human challenges facing the residents of vulnerable communities and the professionals who serve them.

Other federal and state efforts link cancer and community data and make them publicly accessible online. The majority, however, do not provide the ability to parse data at geographic levels smaller than the county. The National Cancer Institute (NCI), for example, offers publicly available interactive tools for mapping and tabulating state—and county-level data on cancer age-adjusted incidence, mortality, screening, and risk factors—filtered by age, sex, race, and ethnicity. These resources draw on data from the National Program of Cancer Registries and Surveillance, Epidemiology, and End Results SEER*Stat Database, U.S. Census, Behavioral Risk Factor Surveillance System (BRFSS), and other public sources [5, 6].

In Florida, the website SCAN 360 [34] (https://www.scan360.com/), a project of Sylvester Comprehensive Cancer Center, serves as a publicly accessible resource for tabulating and visualizing statewide, county-level data on cancer burden, sociodemographics, environmental risk factors, and health behaviors. In Durham County, North Carolina, a restricted set of stakeholders including clinicians and public health practitioners can access the Geographic Health Information System. This innovative system is built from electronic health record (EHR) data from the Duke Health System, vital statistics, property tax, crime, environmental exposure data, and healthcare and community resource data. The Geographic Health Information System is particularly powerful for community health assessment purposes because the Duke system covers the majority of patients in Durham County. In a setting where individual-level clinical data can-not be compiled for a majority of the population, however, a tool such as this may be less useful for community health assessment and surveillance purposes [35].

OH-CASE is an evolving prototype for a more expansive, statewide infrastructure which will integrate new data sources and new features based on stakeholder feedback. An additional goal of this larger effort, which we refer to as CIDaSh (Cancer Informatics and Data Sharing), is to facilitate sharing and linking of research datasets. In the near term, planned data source additions include a statewide database of tobacco outlets and individual-level treatment quality indicators based on Medicare claims and enrollment data [36]. In other states, it is conceivable that available all-payer databases could be integrated into such an infrastructure [37]. In the medium to long term, we plan to add claims data from the state's breast and cervical cancer program (BCCP), disease-specific registry data (e.g., for Lynch Syndrome), air and water quality data, and community risk-factor data. In the coming months, we will add a feature to generate visualizations summarizing query results. Measures visualized will include age-adjusted incidence, proportion of cases presenting with metastatic disease, time-to-treatment, and cancer-specific mortality results; these visualizations will include charts as well as choropleth maps.

Insights and lessons learned

The challenges inherent in a project such as the one described here are many.

- Building the database and interface requires a team of appropriately qualified technical personnel with sufficient bandwidth guided by one or more content experts/champions with a thorough understanding of the source data. It is critical that all parties agree on a plan for providing appropriately detailed written development specifications. Ambiguity in these specifications can cause unnecessary delays and wasted resources.
- Outputs must be validated systematically and on an ongoing basis. We devised a process for randomly selecting queries to be run using the OH-CASE interface with results validated against separate analyses of source data (the topic of a forthcoming manuscript). Briefly, each field and each option contained in the search interface is numbered, and random number generators are used to dictate search criteria for testing. This process has so far helped to identify a major bug occurring when race—and ethnicity-based filter criteria were not

being applied to the denominator used to calculate incidence and mortality rates. Because an analyst must write or modify code to perform each validation and check OH-CASE output against the results of the analysis, validation can be a labor-intensive endeavor.

- Substantial collective thought must go into governance—determining a process for controlling and monitoring access to the data and imposing reasonable obligations for appropriate use. Presently, the OH-CASE Database and Application are overseen by the IRBs of ODH and CWRU. All users must be listed on both IRB protocols and must have completed DUAs and Confidentiality agreements with ODH. Obviously, this arrangement limits the scalability of the user base significantly—especially when it comes to community-based users. We are working with ODH to create a user training and credentialing mechanism, and an authentication mechanism, by which users who are not listed on the IRBs may access the OH-CASE Application (not the underlying Database) based on masking of small numbers in output (any number or combination of numbers revealing a cell size of ten or fewer individuals). Our team is not comfortable with making the OH-CASE interface publicly accessible at any stage in the foreseeable future. The granular nature of the data—even with masking small numbers—requires that users possess a level of understanding that will prevent misinterpretation or inappropriate application of interpreted findings. Users will also likely need some level of access to OH-CASE administrators in order to support appropriate interpretation.
- Sustainability must be a consideration. After the setup phase—even if new data types are not subsequently added—periodic data updates and validation will be required. In addition, a process of granting and controlling user access must be maintained. Ideally, the needed resources will come from institutional support, meaning that champions will need to demonstrate a value proposition to institutional leadership. The mandate to show catchment area insight and impact can be a strong motivator for National Cancer Institute (NCI) designated Clinical or Comprehensive Cancer Centers [38].

To plan for these challenges, we have so far sought the input of the eight-member advisory committee serving the cancer center shared resource where OH-CASE is based. Members consist of PhD researchers and clinician scientists affiliated with Case Comprehensive Cancer Center. As we begin to broaden access to the OH-CASE Application, we will convene an OH-CASE steering committee which includes a subset of this institutional group as well as members from community groups likely to use the application. The steering committee will guide us in matters of governance as well as in prioritizing additional data sources and application features.

Conclusion

By broadening the user base and lowering barriers to usage, the model we have described here has the potential to increase the value of available cancer and community data, enhance the relevance of community-partnered research, and improve the effectiveness of cancer control efforts at the local level. These benefits can accrue through the synergies created when academic data stewards and methodologists include community-based practitioners in the full life cycle of data linkage efforts.

Acknowledgments

This study includes data provided by the Ohio Department of Health which should not be considered an endorsement of this study or its conclusions. We would like to thank Lynn Giljahn and Roberta Slocumb at the Ohio Department of Health for their invaluable assistance.

Funding

Development of the work described here has been supported by funding from the National Cancer Institute, Case Comprehensive Cancer Center (P30 CA043703) and from the Cleveland Translational Science Collaborative (UL1TR002548), the Case Western Reserve University Center for Reducing Health Disparities (U54MD002265), and Susan G. Komen of Northeast Ohio. None of these supporters played a role in acquiring data, building the data resource described, writing the manuscript, or the decision to submit the article for publication.

Data availability

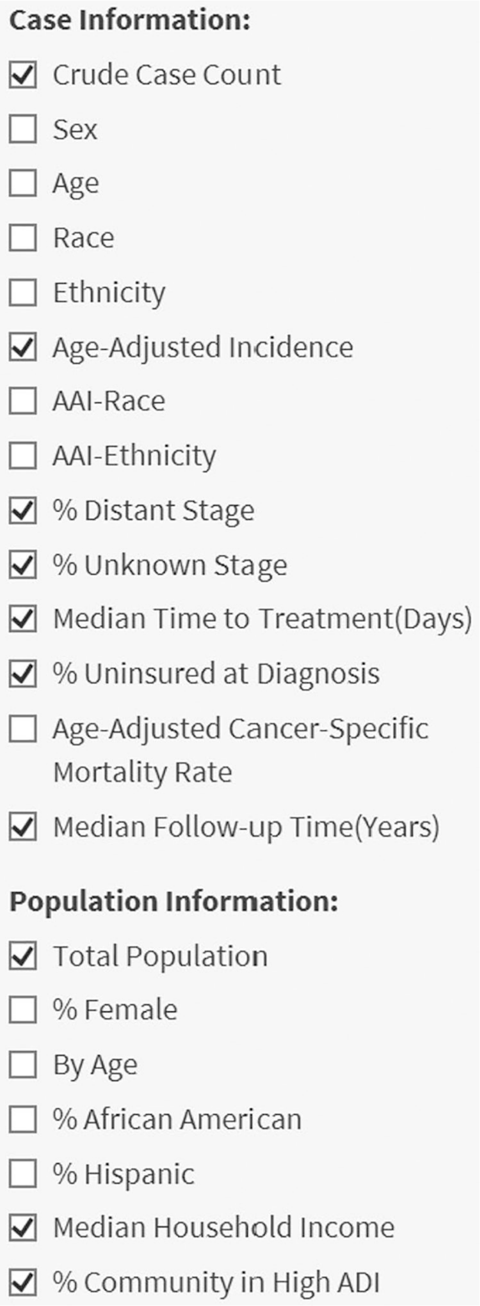
All data sources cited in the study are publicly available except for individual-level data from the Ohio Cancer Incidence Surveillance System (OCISS), which is used under confidentiality and data use agreement with the Ohio Department of Health.

Appendix

Case study 1: analysis of racial disparities in cancer incidence and mortality across three neighboring counties

Any of the following stakeholders may be interested in a broad comparison of disparities in age-adjusted cancer incidence and mortality across local communities:

- Cancer center leadership seeking to understand disparities in cancer burden within their catchment area
- Researchers attempting to narrow their sampling frame for a trial of a multi-level intervention to reduce community cancer disparities
- A health improvement planning collaborative seeking to understand the relative extent of racial disparities in cancer burden within their jurisdiction



Case Information:

☑ Crude Case Count

□ Sex

□ Age

□ Race

□ Ethnicity

☑ Age-Adjusted Incidence

□ AAI-Race

□ AAI-Ethnicity

☑ % Distant Stage

☑ % Unknown Stage

☑ Median Time to Treatment(Days)

☑ % Uninsured at Diagnosis

□ Age-Adjusted Cancer-Specific Mortality Rate

☑ Median Follow-up Time(Years)

Population Information:

☑ Total Population

□ % Female

□ By Age

□ % African American

□ % Hispanic

☑ Median Household Income

☑ % Community in High ADI

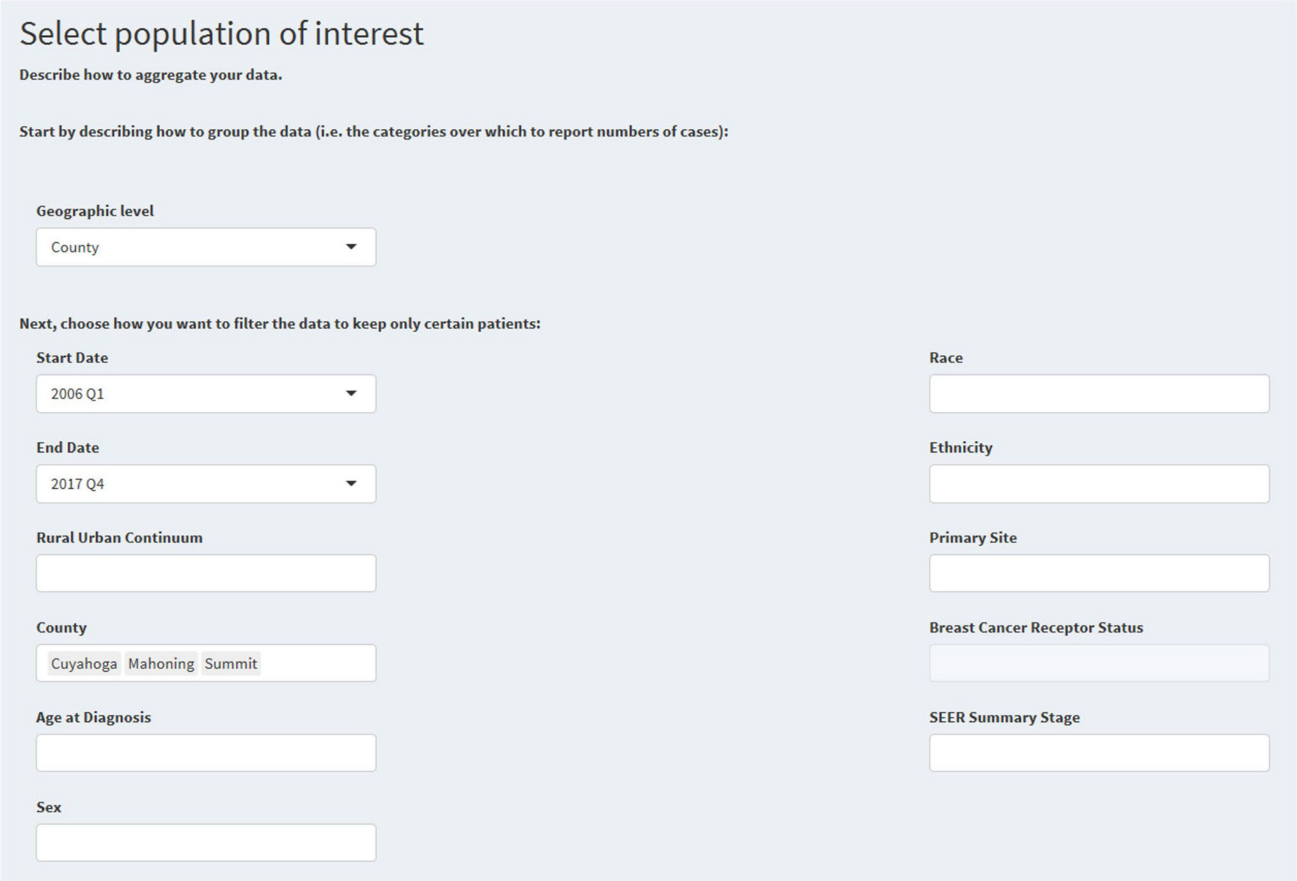
Fig. 3.

OH-CASE Column display options—The boxes checked here are displayed by default.

Figure 4 shows search criteria for a query comparing three neighboring counties in northeast Ohio. Figure 5 shows relevant data fields from query results.

Interpretation and insights—For both race groups, substantial differences in age-adjusted incidence existed over the period of interest between Cuyahoga County and neighboring Summit and Mahoning counties. Mortality was higher in all counties for AA’s compared to non-AAs, but racial mortality disparities were most pronounced in Cuyahoga and Mahoning.

An analysis such as this can serve as the starting point for decisions on where to focus attention, resources, or data gathering for numerous stakeholders.



Select population of interest

Describe how to aggregate your data.

Start by describing how to group the data (i.e. the categories over which to report numbers of cases):

Geographic level: County

Next, choose how you want to filter the data to keep only certain patients:

Start Date: 2006 Q1

End Date: 2017 Q4

Rural Urban Continuum:

County: Cuyahoga Mahoning Summit

Age at Diagnosis:

Sex:

Race:

Ethnicity:

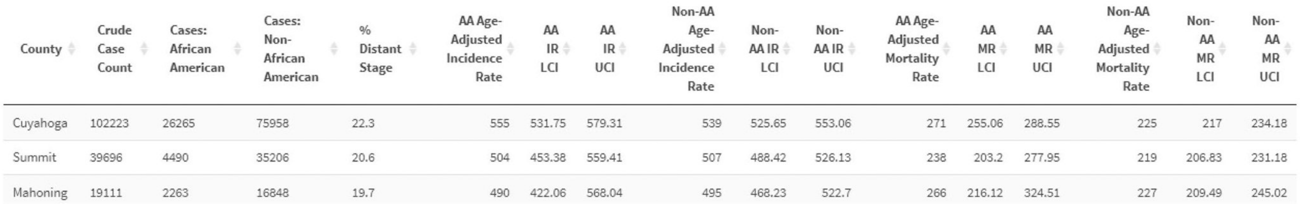
Primary Site:

Breast Cancer Receptor Status:

SEER Summary Stage:

Fig. 4.

Search criteria for a county-level query of all invasive cancer cases from 2006 to 2017 in a three county area of northeast Ohio



County	Crude Case Count	Cases: African American	Cases: Non-African American	% Distant Stage	AA Age-Adjusted Incidence Rate	AA IR LCI	AA IR UCI	Non-AA Age-Adjusted Incidence Rate	Non-AA IR LCI	Non-AA IR UCI	AA Age-Adjusted Mortality Rate	AA MR LCI	AA MR UCI	Non-AA Age-Adjusted Mortality Rate	Non-AA MR LCI	Non-AA MR UCI
Cuyahoga	102223	26265	75958	22.3	555	531.75	579.31	539	525.65	553.06	271	255.06	288.55	225	217	234.18
Summit	39696	4490	35206	20.6	504	453.38	559.41	507	488.42	526.13	238	203.2	277.95	219	206.83	231.18
Mahoning	19111	2263	16848	19.7	490	422.06	568.04	495	468.23	522.7	266	216.12	324.51	227	209.49	245.02

Fig. 5.

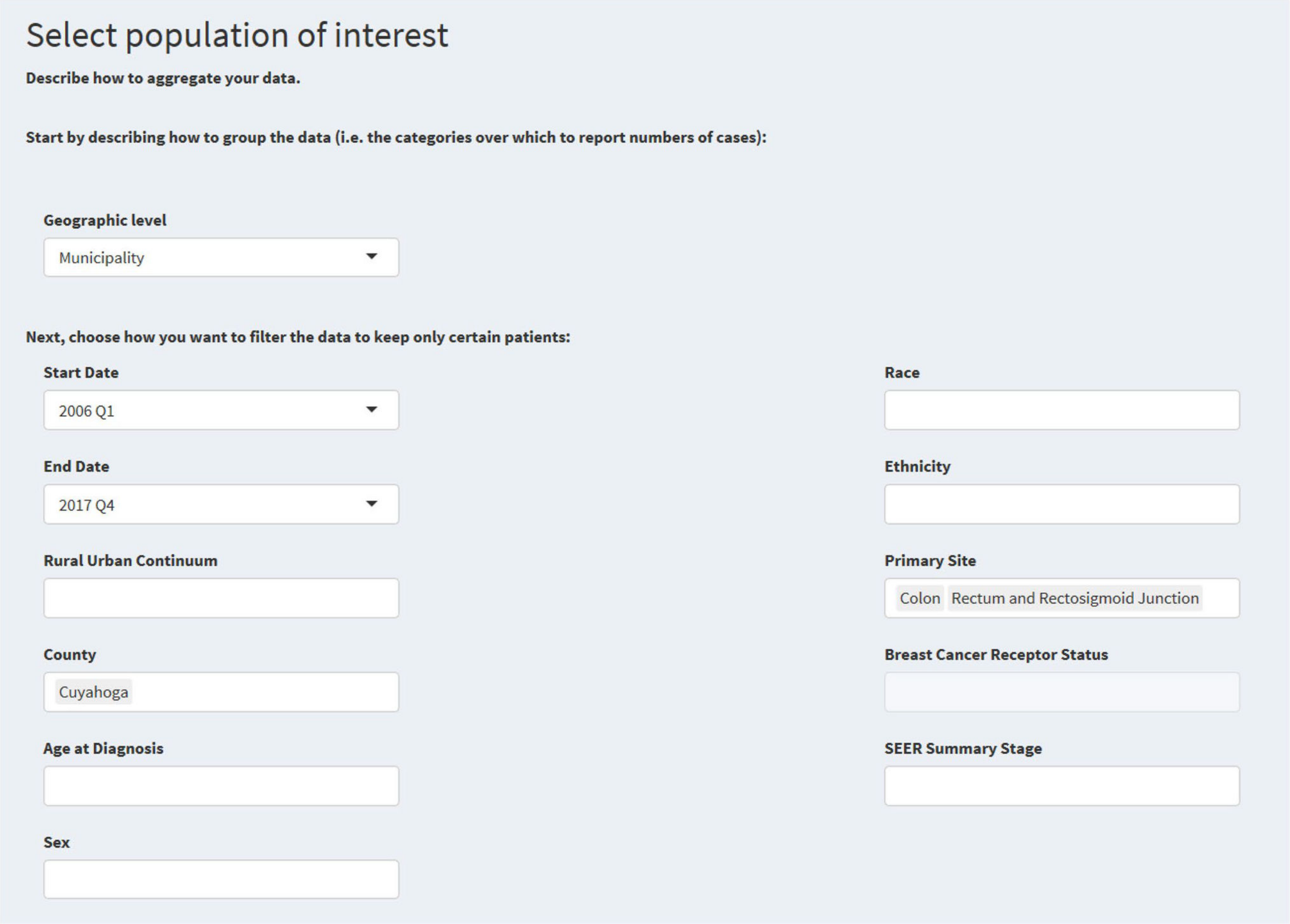
Selected columns from query results for the search specified in Fig. 4; AA = African American, IR = Incidence Rate, LCI = Lower confidence interval, UCI = Upper confidence interval

Case study 2: comparing colorectal cancer outcomes across municipalities within an urban county

Any of the following stakeholders may be interested in a comparison of colorectal cancer (CRC) outcomes across communities within a single county.

- A county public health department making decisions about which interventions to adopt and where to deploy resources from a grant targeted at boosting primary and secondary CRC prevention
- An implementation researcher trying to understand what characteristics or resources can lead to divergent cancer outcomes in similar communities
- A local health system making programmatic decisions about where to focus efforts to grow “Flu-FIT” programs which distribute fecal immunochemical testing (FIT) kits at the time of flu shot administration

Figure 6 shows search criteria for a query comparing municipalities within Cuyahoga County, Ohio. Figure 7 shows relevant data fields from query results.



Select population of interest

Describe how to aggregate your data.

Start by describing how to group the data (i.e. the categories over which to report numbers of cases):

Geographic level: Municipality

Next, choose how you want to filter the data to keep only certain patients:

Start Date: 2006 Q1

End Date: 2017 Q4

Rural Urban Continuum:

County: Cuyahoga

Age at Diagnosis:

Sex:

Race:

Ethnicity:

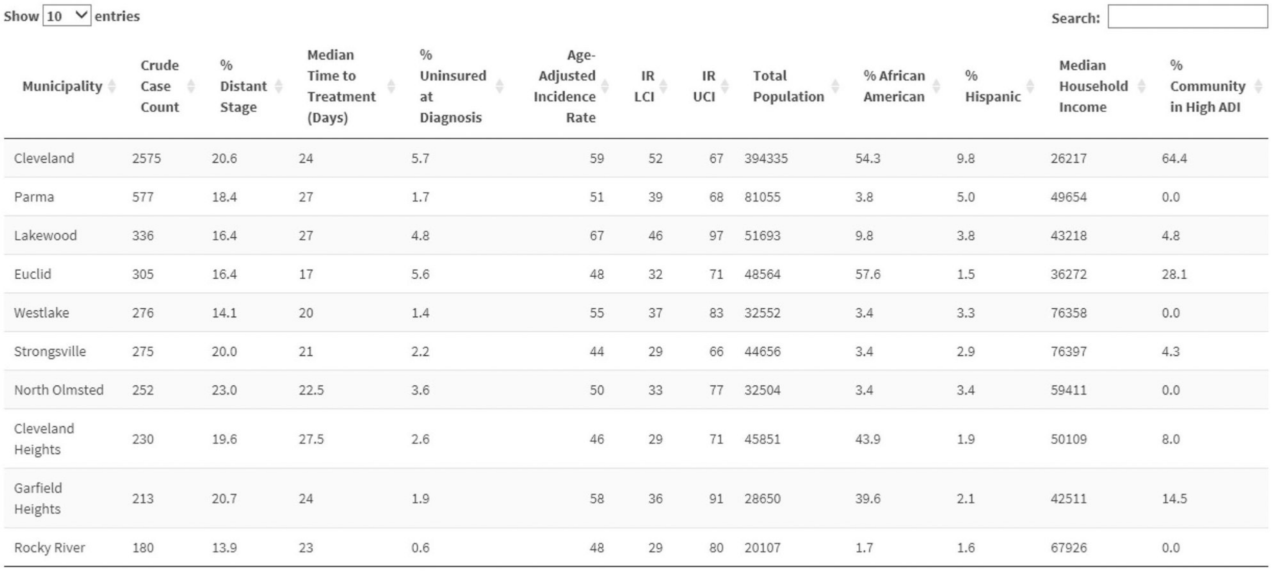
Primary Site: Colon, Rectum and Rectosigmoid Junction

Breast Cancer Receptor Status:

SEER Summary Stage:

Fig. 6.

Search criteria for a municipality-level search of all colorectal cancer cases from 2006 to 2017 in Cuyahoga County, Ohio



Municipality	Crude Case Count	% Distant Stage	Median Time to Treatment (Days)	% Uninsured at Diagnosis	Age-Adjusted Incidence Rate	IR LCI	IR UCI	Total Population	% African American	% Hispanic	Median Household Income	% Community in High ADI
Cleveland	2575	20.6	24	5.7	59	52	67	394335	54.3	9.8	26217	64.4
Parma	577	18.4	27	1.7	51	39	68	81055	3.8	5.0	49654	0.0
Lakewood	336	16.4	27	4.8	67	46	97	51693	9.8	3.8	43218	4.8
Euclid	305	16.4	17	5.6	48	32	71	48564	57.6	1.5	36272	28.1
Westlake	276	14.1	20	1.4	55	37	83	32552	3.4	3.3	76358	0.0
Strongsville	275	20.0	21	2.2	44	29	66	44656	3.4	2.9	76397	4.3
North Olmsted	252	23.0	22.5	3.6	50	33	77	32504	3.4	3.4	59411	0.0
Cleveland Heights	230	19.6	27.5	2.6	46	29	71	45851	43.9	1.9	50109	8.0
Garfield Heights	213	20.7	24	1.9	58	36	91	28650	39.6	2.1	42511	14.5
Rocky River	180	13.9	23	0.6	48	29	80	20107	1.7	1.6	67926	0.0

Fig. 7.
Selected columns from query results for the search specified in Fig. 6; IR = Incidence Rate, LCI = Lower confidence interval, UCI = Upper confidence interval, ADI = Area Deprivation index [24]

Interpretation and insights—A number of potential insights arise from the output shown in Fig. 7. Some clear dichotomies exist between pairs of similar communities which could inform the actions of the county health department, the implementation researcher, or the health system described above. Consider the communities of Westlake and North Olmstead. They are relatively affluent, predominantly white, similar in size, and located on the suburban west side of Cleveland. They have comparable CRC incidence, but substantially different stage distributions. The proportion of metastatic disease is over nine percentage points higher (over 60% in relative terms) in North Olmstead. This difference warrants investigation and may indicate a need for efforts to improve screening uptake in North Olmstead.

Secondly, examining the case of Euclid Ohio may yield important lessons for researchers, public health officials, advocacy groups, and others. It is an economically diverse community with modest median income and a high proportion of individuals living in resource-deprived neighborhoods, and it has a high proportion of uninsured individuals. Nonetheless, Euclid performs as well as or better than most of its neighbors shown in Fig. 7—including those with significantly more resources—in terms of cancer incidence, metastatic disease, and time-to-treatment. Understanding what is going *right* in Euclid, building on it, and emulating it could yield benefits within and beyond Cuyahoga County.

References

1. Polite BN, Adams-Campbell LL, Brawley OW, Bickell N, Care-thers JM, Flowers CR, Foti M, Gomez SL, Griggs JJ, Lathan CS, Li CI, Lichtenfeld JL, McCaskill-Stevens W, Paskett ED (2017) Charting the future of cancer health disparities research: a position statement from the American Association for Cancer Research, the American Cancer Society, the American Society of Clinical Oncology, and the National Cancer Institute. CA Cancer J Clin. 10.3322/caac.21404
2. Gehlert S, Coleman R (2010) Using community-based participatory research to ameliorate cancer disparities. Heal Soc Work 35:304–309. 10.1093/hsw/35.4.302
3. Vargas R, Maxwell AE, Lucas-Wright A, Bazargan M, Barlett C, Jones F, Brown A, Forge N, Smith J, Vadgamma J, Jones L (2014) A community partnered-participatory research approach to reduce cancer disparities in South Los Angeles. Prog Commun Heal Partnerships Res Educ Action 8:471–476. 10.1353/cpr.2014.0063
4. Edmunds M, Thorpe L, Sepulveda M, Bezold C, Ross DA (2014) The future of public health informatics: alternative scenarios and recommended strategies. EGEMs 2:1156. 10.13063/2327-9214.1156 [PubMed: 25848630]
5. National Cancer Institute, State Cancer Profiles (2020) statecancerprofiles.cancer.gov.
6. National Cancer Institute, SEER*Explorer (2021). https://seer.cancer.gov/explorer/. Accessed 3 Jan 2021
7. Sylvester Comprehensive Cancer Center, Scan 360 (2020). https://www.scan360.com/
8. Centers for Disease Control and Prevention, National Program of Cancer Registries (2021). https://www.cdc.gov/cancer/npcr/index.htm. Accessed 3 Jan 2021
9. North American Association of Cancer Center Registries, Cancer Registry Standards (2018). https://www.naaccr.org/
10. United States Census, About the American Community Survey (2021). https://www.census.gov/programs-surveys/acs/about.html. Accessed 2 Jan 2021.
11. Dalzell LP, Tangka FKL, Powers DS, O'Hara BJ, Holmes W, Joseph K, Royalty J (2015) Data sources for identifying low-income, uninsured populations: application to public health—National Breast and Cervical Cancer Early Detection Program. Cancer Causes Control 26:699–709. 10.1007/s10552-015-0571-y [PubMed: 25916228]
12. Federal Register, Fed. Regist. 83 FR 5629 (2018) 56293–56298. https://www.federalregister.gov/d/2018-24570
13. Health Resources and Services Administration, What is shortage designation (2021). https://bhw.hrsa.gov/shortage-designation/hpsas
14. U.S. Food and Drug Administration, Certified Facilities/Certificates, (2020). https://bhw.hrsa.gov/shortage-designation/hpsas.
15. Koroukian SM, Cooper GS, Rimm AA (2003) Ability of Medicaid claims data to identify incident cases of breast cancer in the Ohio Medicaid population. Health Serv Res 38:947–960. 10.1111/1475-6773.00155 [PubMed: 12822920]
16. World Health Organization (2008) international classification of diseases for oncology, 3rd edn. WHO, Geneva
17. National Cancer Institute, SEER Program and Coding Manual (2021). https://seer.cancer.gov/tools/codingmanuals/
18. American Joint Committee on Cancer (2018) AJCC Cancer Staging Manual, 8th ed. https://cancerstaging.org
19. Curtin LR, Klein RJ (1955) Direct Standardization (Age-Adjusted Death Rates), Stat. Notes
20. N.C. Institute (2020) Standard Populations (Millions) for Age-Adjustment. https://seer.cancer.gov/stdpopulations/
21. Koroukian SM, Xu F, Beaird H, Diaz M, Murray P, Rose JH (2007) Complexity of care needs and unstaged cancer in elders: a population-based study. Cancer Detect Prev 31:199–206. 10.1016/j.cdp.2007.04.002 [PubMed: 17658225]
22. Koroukian SM (2003) Assessing the effectiveness of medicaid in breast and cervical cancer prevention. J Health Manag Pract. 2003:306–314

23. Hanna TP, King WD, Thibodeau S, Jalink M, Paulin GA, Harvey-jones E, Sullivan DEO, Booth CM, Sullivan R, Aggarwal A (2020) Mortality due to cancer treatment delay: systematic review. BMJ 4(371):m4087. 10.1136/bmj.m4087

24. Singh GK (2003) Area deprivation and widening inequalities in US Mortality, 1969–1998. Am J Pub Heal 93:1137–1143

25. Kim U, Koroukian S, Statler A, Rose J (2020) The effect of medicaid expansion among adults from low-income communities on stage at diagnosis in those with screening-amenable cancers. Cancer 126:4209–4219. 10.1002/cncr.32895 [PubMed: 32627180]

26. Rose J, Kim U, Dong W, Hnath J, Koopman S, Oliver Y, Sage P, Koroukian S. A qualitative study of factors affecting timely breast cancer treatment in a high-risk Cleveland Community, Case Comprehensive Cancer Century 4th Annual Disparities Symposium (n.d.)

27. Rose J, Dong W, Kim U, Obeng-Gyasi S, Koroukian S. Medicaid expansion associated with earlier stage and improved reconstruction rates in low income breast cancer patients. In: San Antonio Breast Cancer Symposium, Virtual, n.d.

28. Parrott R, Volkman JE, Lengerich E, Ghetian CB, Chadwick AE, Hopfer S (2010) Using geographic information systems to promote community involvement in comprehensive cancer control. Health Commun 25:276–285. 10.1080/10410231003711755 [PubMed: 20461613]

29. Chin MH, Walters AE, Cook SC, Huang ES (2007) Interventions to reduce racial and ethnic disparities in health care. Med Care Res Rev 64:7S–28S. 10.1177/1077558707305413 [PubMed: 17881624]

30. Meade CD, Menard JM, Luque JS, Martinez-Tyson D, Gwede CK (2011) Creating community-academic partnerships for cancer disparities research and health promotion. Health Promot Pract 12:456–462. 10.1177/1524839909341035 [PubMed: 19822724]

31. Corbin JH, Fernandez ME, Mullen PD (2015) Evaluation of a community-academic partnership: lessons from Latinos in a network for cancer control. Health Promot Pract 16:345–353. 10.1177/1524839914558514 [PubMed: 25395057]

32. Wynn TA, Anderson-Lewis C, Johnson R, Hardy C, Hardin G, Walker S, Marron J, Fouad M, Partridge E, Scarinci I (2011) Developing a community action plan to eliminate cancer disparities: lessons learned. Prog Community Health Partnersh 5:161–168. 10.1353/cpr.2011.0013 [PubMed: 21623018]

33. Noel L, Phillips F, Tossas-milligan K, Spear K, Vanderford NL, Eckhardt SG (2019) Community-academic partnerships: approaches to engagement, 2019 ASCO Education B

34. Scan360 (2019) https://www.scan360.com/. Accessed 29 Dec 2019

35. Miranda ML, Ferranti J, Strauss B, Neelon B, Califf RM (2013) Geographic health information systems: a platform to support the “triple aim”. Health Aff 32:1608–1615. 10.1377/hlthaff.2012.1199

36. Tucker TC, Durbin EB, McDowell JK, Huang B (2019) Unlocking the potential of population-based cancer registries. Cancer 125:3729–3737. 10.1002/cncr.32355 [PubMed: 31381143]

37. National Conference of State Legislators, Collecting health data: all-payer claims databases (2019). http://www.ncsl.org/research/health/collecting-health-data-all-payer-claims-database.aspx. Accessed 29 Dec 2019

38. Department of Health and Human Services, Founding Opportunity Announcement: Cancer Center Support Grants (CCSGs) for NCI-designated Cancer Centers (P30 Clinical Trial Optional) (2019). https://grants.nih.gov/grants/guide/pa-files/par-20-043.html. Accessed 14 Apr 2021

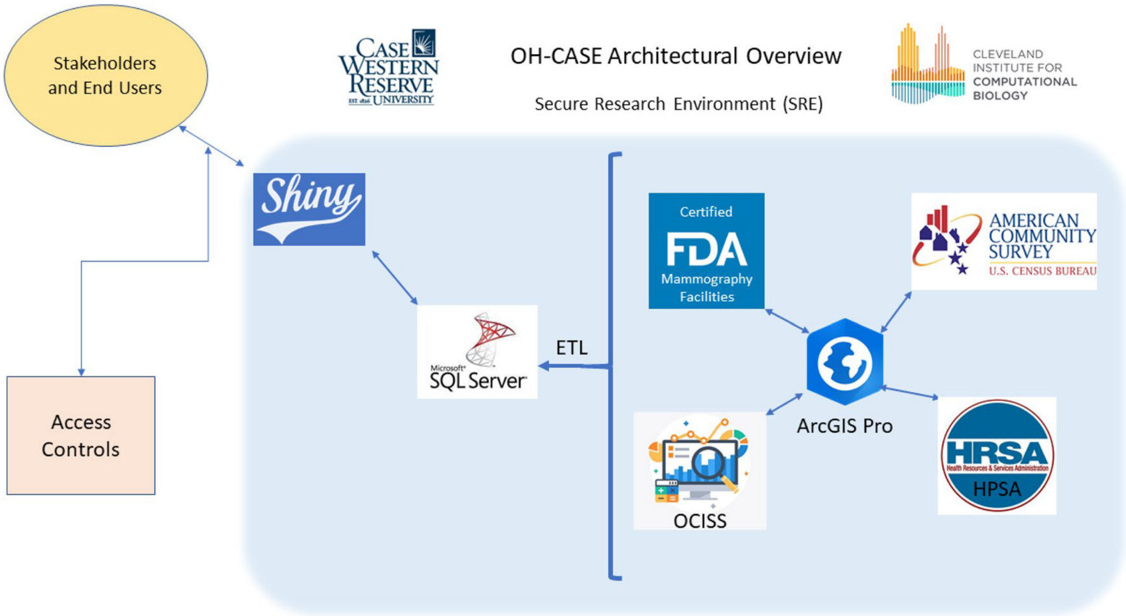
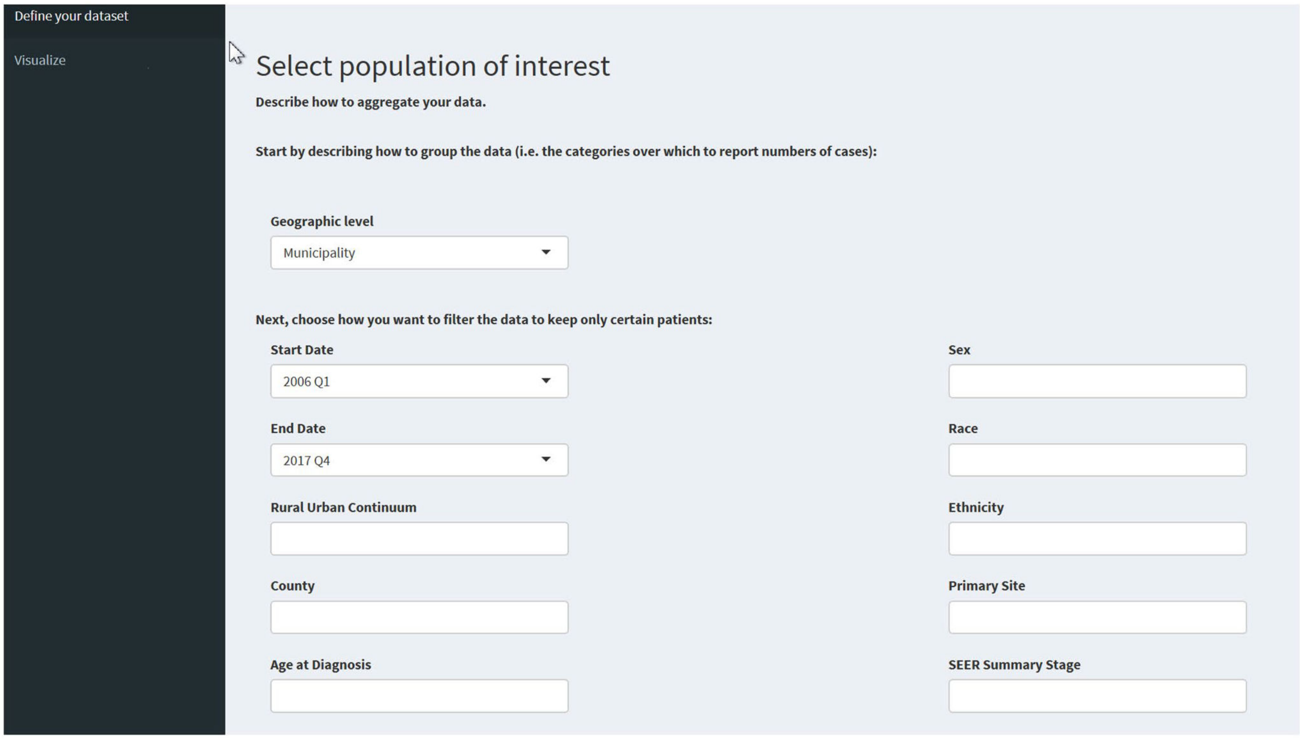

OH-CASE Architectural Overview
Secure Research Environment (SRE)
CASE WESTERN RESERVE UNIVERSITY
CLEVELAND INSTITUTE FOR COMPUTATIONAL BIOLOGY
Stakeholders and End Users
Access Controls
Shiny
Microsoft SQL Server
ETL
Certified FDA Mammography Facilities
AMERICAN COMMUNITY SURVEY U.S. CENSUS BUREAU
ArcGIS Pro
OCISS
HRSA HPSA

Fig. 1.

OH-CASE Architecture—FDA = U.S. Food and Drug Administration; OCISS = Ohio Cancer Incidence Surveillance System; HRSA = Health Resources and Services Administration; ETL = extraction, transformation, and loading



Select population of interest

Describe how to aggregate your data.

Start by describing how to group the data (i.e. the categories over which to report numbers of cases):

Geographic level: Municipality

Next, choose how you want to filter the data to keep only certain patients:

Start Date: 2006 Q1

End Date: 2017 Q4

Rural Urban Continuum:

County:

Age at Diagnosis:

Sex:

Race:

Ethnicity:

Primary Site:

SEER Summary Stage:

Fig. 2.

OH-CASE interface filter criteria fields

Table 1

Incident Invasive cancer cases from 2014 to 2018 in Ohio by race, ethnicity, sex, and cancer site/type

	2014–2018 Incident Invasive Cancer Cases (%)	
All cancer sites/types	337,292	(100.0)
Age [mean/median]	65.3/66	
Race		
American Indian/Alaska Native	219	(0.1)
Asian	2,471	(0.7)
Black	33,915	(10.1)
Native Hawaiian/Pacific Islander	137	(0.0)
White	294,720	(87.4)
Other	1909	(0.6)
Unknown	3,921	(1.2)
Ethnicity		
Hispanic	2,908	(0.9)
Non-Hispanic	324,541	(96.2)
Unknown Ethnicity	9,843	(2.9)
Sex		
Female	168,334	(50.0)
Male	168,939	(50.0)
Unknown/Other	19	(0.0)
Cancer site/type		
Lung and Bronchus	50,293	(14.9)
Breast (Female)	48,516	(14.4)
Prostate	38,769	(11.5)
Colorectal	29,573	(8.8)
Melanoma	17,015	(5.0)
Bladder	16,122	(4.8)
Non-Hodgkin’s Lymphoma	13,664	(4.1)
Kidney and Renal Pelvis	12,516	(3.7)
Uterine	12,255	(3.6)
Thyroid	9,346	(2.8)
All other	89,223	(26.5)

Table 2

Use case examples compiled based on stakeholder input

Stakeholder Group	Use Cases
Cancer center community outreach teams	• Prioritize communities needing services from a mobile mammography unit based on high rates of late-stage breast cancer diagnosis, low mammography facility density, and poor transportation access • Compare certain process measures (e.g., time from diagnosis to treatment), risk factors, and outcomes (e.g., incidence, mortality, stage at diagnosis) for patients across communities, benchmarked against state/national averages • Report outcomes for NCI-designated cancer center catchment areas
Cancer advocacy organizations	• Prioritize communities/subpopulations for advocacy and funding activities • Compile descriptive cancer burden statistics for a population of particular interest to a foundation grant funder • Provide data to inform government-targeted advocacy activities
Local public health	• Support applications for state and federal funding with cancer burden and demographic statistics • Support cancer cluster investigations stemming from citizen inquiries
Researchers	• Provide community-level data to local partners to stimulate hypothesis generation and inform study design as a component of community-engaged/partnered research • Identify the size and demographic makeup of relevant sampling frames as part of determining study feasibility • Compile inclusion enrollment data for population-based study proposals • Use demographic, process, and outcomes data to target primary data collection (e.g., semi-structured interviews) in disparities research