Review Article

Unlocking the Potential of Population-Based Cancer Registries

Thomas C. Tucker, PhD, MPH[1,2]; Eric B. Durbin, DrPH, MS[1,3]; Jaclyn K. McDowell, DrPH[1,2]; and Bin Huang, MS, DrPH[1,4]

Population-based cancer registries have improved dramatically over the last 2 decades. These central cancer registries provide a critical framework that can elevate the science of cancer research. There have also been important technical and scientific advances that help to unlock the potential of population-based cancer registries. These advances include improvements in probabilistic record linkage, refinements in natural language processing, the ability to perform genomic sequencing on formalin-fixed, paraffin-embedded (FFPE) tissue, and improvements in the ability to identify activity levels of many different signaling molecules in FFPE tissue. This article describes how central cancer registries can provide a population-based sample frame that will lead to studies with strong external validity, how central cancer registries can link with public and private health insurance claims to obtain complete treatment information, how central cancer registries can use informatics techniques to provide population-based rapid case ascertainment, how central cancer registries can serve as a population-based virtual tissue repository, and how population-based cancer registries are essential for guiding the implementation of evidence-based interventions and measuring changes in the cancer burden after the implementation of these interventions. ***Cancer* 2019;125:3729-3737**.

KEYWORDS: cancer burden, cancer surveillance, central cancer registries, external validity, outcomes research, population-based.

INTRODUCTION

Over the past several decades, population-based cancer registries have improved dramatically in both quantity and quality. There are 2 separate federal programs that provide financial support for population-based cancer surveillance programs in the United States. These programs are the Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute (NCI) and the National Program of Cancer Registries (NPCR) of the Centers for Disease Control and Prevention (CDC).[1]

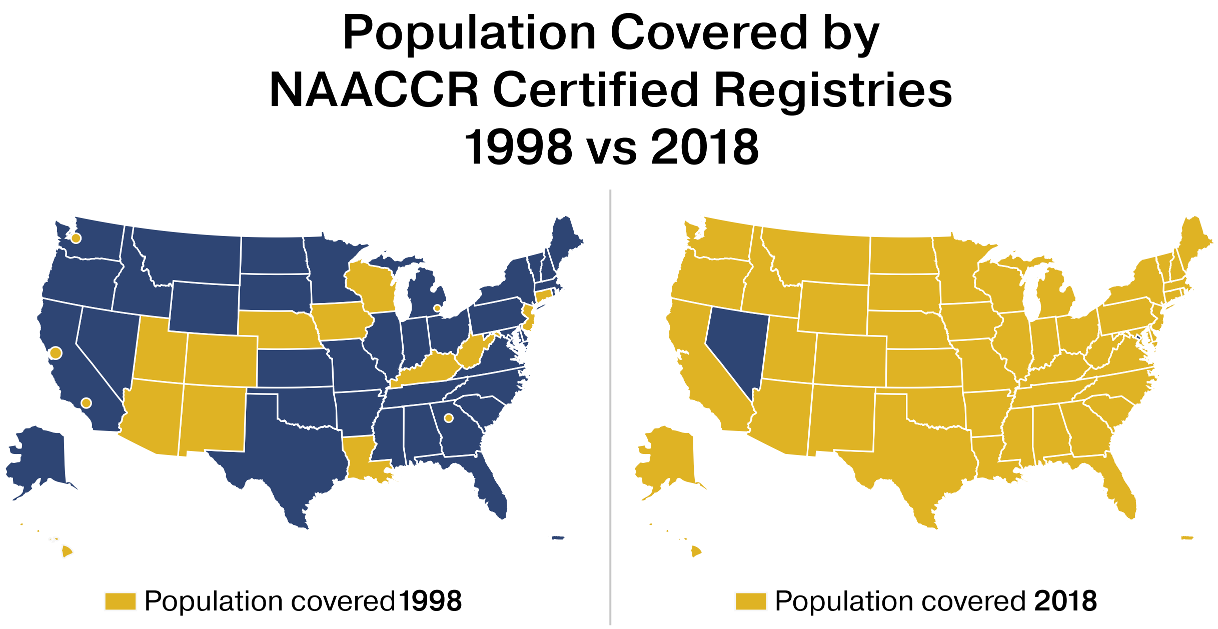
The NCI SEER program was established in 1973 to address components of the 1971 National Cancer Act.[1] The NCI SEER program began with 7 population-based cancer registries. Over the years, the SEER program has expanded several times and now includes 20 state and urban population–based cancer registries.[2] The SEER program works to "provide information on cancer statistics in an effort to reduce the burden of cancer among the U.S. population."[3] The CDC NPCR was established in 1992 by the Cancer Registries Amendment Act. Before the NPCR, "10 states had no registry, and most states with registries lacked the resources and legislative support they needed to gather complete data."[4] Today, the NPCR supports population-based cancer registries in 46 US states, the District of Columbia, Puerto Rico, the US Pacific Island Jurisdictions, and the US Virgin Islands.[4] Together, the SEER and NPCR programs provide support for central cancer registries that cover the populations of all US states and the District of Columbia as well as US territories.

Along with the increased number of population-based cancer registries, there has been a substantial increase in both the quantity and quality of data collected by these surveillance programs. In 1995, the North American Association of Central Cancer Registries (NAACCR) brought together representatives from all of the cancer surveillance standard setting organizations in the United States and Canada to form the Uniform Data Standards Committee. This committee annually publishes a dictionary of all of the data variables (and their definitions) that population-based cancer registries are required to collect. In 1997, there were 253 required variables.[5] Since that time, the manner in which specific cancers are diagnosed and treated has changed. Even the way some cancers are defined has changed.[6,7] Today, modern population-based cancer registries collect 785 required data variables.[8]

Corresponding author: Thomas C. Tucker, PhD, MPH, Markey Cancer Center, University of Kentucky, 2365 Harrodsburg Road, Suite A230, Lexington, KY 40504; tct@kcr.uky.edu

[1]Kentucky Cancer Registry, Markey Cancer Center, University of Kentucky, Lexington, Kentucky; [2]Department of Epidemiology, College of Public Health, University of Kentucky, Lexington, Kentucky; [3]Division of Biomedical Informatics, Department of Internal Medicine, College of Medicine, University of Kentucky, Lexington, Kentucky; [4]Department of Biostatistics, College of Public Health, University of Kentucky, Lexington, Kentucky

DOI: 10.1002/cncr.32355, **Received:** August 8, 2018; **Revised:** March 13, 2019; **Accepted:** April 16, 2019, **Published online** August 5, 2019 in Wiley Online Library (wileyonlinelibrary.com)



Source: North American Association of Central Cancer Registries: Certified Registries.

Figure 1. US populations covered by NAACCR-certified registries: 1998 versus 2018.

In 1997, NAACCR established criteria for the certification of population-based cancer registries. The criteria are designed to measure the completeness, accuracy, and timeliness of data collected by each registry.[9] Annually, these objective criteria are independently applied to the data from each population-based cancer registry in North America. During the first year in which the criteria were implemented, only 17 US population-based cancer registries met the criteria. Today, all but 2 central cancer registries in the United States (n = 56) meet these quality criteria[10] (see Fig. 1).

There have also been a number of important technical and scientific advances that have the potential to greatly enhance the utility of population-based cancer registries and can significantly improve cancer research. These advances include improvements in probabilistic record linkage, refinements in natural language processing (NLP), the ability to perform genomic sequencing on formalin-fixed, paraffin-embedded (FFPE) tissue, and improvements in the ability to identify activity levels of many different signaling molecules in FFPE tissue via immunohistochemical staining techniques.[11-19]

This article describes how central cancer registries can be used to improve the science in cancer research by providing a population-based sample frame that will lead to studies with strong external validity. We explore how advances in informatics and laboratory science make it possible for central cancer registries to obtain complete treatment information, provide population-based rapid case ascertainment, and serve as a population-based virtual tissue repository (VTR). We also describe the important role of population-based cancer registries in guiding the implementation of evidence-based interventions and measuring changes in the population after the implementation of these interventions.

USING THE POPULATION-BASED CANCER REGISTRY AS A SAMPLE FRAME

One of the challenges for cancer researchers is how to obtain a random sample of cases that scientifically represent the underlying population. In other words, how can a set of cases be obtained that will have strong external validity and will allow investigators to generalize the findings from their studies to some underlying population? One example of the bias that can result from not having a true random sample with external validity is as follows: If an investigator wants to identify geographic variations in the treatment patterns for a specific cancer and uses data only from hospitals with multidisciplinary cancer programs to explore this issue, the results will most likely be biased because hospitals without a multidisciplinary cancer program will be systematically excluded. Even with a large sample size, the bias will remain because the records in this hypothetical study will always exclude patients treated in hospitals without multidisciplinary cancer programs. These hospitals are often smaller, are more likely to be in rural areas, and frequently treat patients who have fewer financial resources.[20] Consequently, studies that do not use a sample that truly represents the underlying population will be limited by selection bias and lack external validity.

Population-based cancer registries make it possible for investigators to obtain a set of case records that truly represent the entire underlying population. These surveillance systems collect detailed data on every new case of cancer diagnosed in the geographic area covered by the registry. The data are collected with uniform data standards and definitions.[8] Thus, the meaning of each variable in the data set is the same throughout North America.[15] The data are subjected to independent, objective quality control measures regarding completeness, accuracy, and timeliness.[8,9] Because data are collected on all new cases of cancer that occur in the population covered by the registry, the records collected by the central registry create a population-based sample frame that can be used to provide researchers with a set of records that truly represent the cancer cases occurring from the underlying population. As a result, studies using these data sets will have strong external validity. There are many examples of how population-based cancer registries have been used to conduct studies that have strong external validity.[21-24]

Another challenge for investigators conducting population-based cancer outcomes research has been obtaining a data set with complete treatment information. Patients with cancer are frequently treated at multiple facilities. They can receive surgery in a hospital, radiation therapy in a private freestanding outpatient facility, and chemotherapy in a medical oncology practice office. These facilities can be in widely disparate locations, even in different states. It is also important to note that treatment for cancer can take place over many months or even years. This makes it difficult for any cancer surveillance program to collect all of the relevant treatment information on all of the cases occurring in a population. The emergence and refinement of probabilistic record linkage have helped to alleviate this problem. Probabilistic record linkage is a method commonly used to determine whether demographic records refer to the same person.[25]

For nearly 2 decades, SEER population-based cancer registry records have been linked with Medicare health insurance claims to obtain complete treatment information for patients with cancer who are 65 years old or older. These record linkage efforts have shown that, overall, 32% of cancer cases aged 65 years or older were missing chemotherapy treatment data, and 20% of the cases aged 65 years or older were missing radiation therapy data.[26] Linking the records from a population-based cancer registry with Medicare claims is believed to provide nearly complete treatment information for patients 65 years old or older.

Patients with cancer who are younger than 65 years are still very likely missing treatment data. Consequently, population-based outcomes research studies for determining disparities in cancer treatment that use data from central cancer registries have historically been limited to cancer cases aged 65 years or older that have been linked with Medicare claims.[27,28] To obtain all of the missing treatment information for patients 65 years and older as well as those younger than 65 years requires linking the cancer registry records with all public and private health insurance claims.[29,30] This has already been successfully accomplished by a number of population-based cancer registries.[31-35] The establishment of all-payer claims databases in 26 states has helped to facilitate these efforts.[36] Even in states that do not have an all-payer claims database, population-based cancer registries have been successful in linking with the health insurance claims from nearly all of the health insurance companies in the state.[34] As part of these record linkage efforts, algorithms have been created to extract the relevant treatment information from the claims data, including genomic testing, and this additional information has been used to supplement the treatment records in the central cancer registry. In addition, using health insurance claims data 1 year before diagnosis, some of the record linkage efforts have constructed comorbidity indices for each patient with cancer.[37]

The efforts to develop a system to link the cancer registry records with all of the public and private health insurance claims are not trivial. However, a number of population-based cancer registries have demonstrated that it is possible to establish such a linkage and create a population-based research data set with nearly complete treatment information for patients with cancer of all ages.[28,31,33,34,38-40]

USING THE POPULATION-BASED CANCER REGISTRY FOR RAPID CASE ASCERTAINMENT

Central cancer registries have been criticized for not being able to provide population-based data rapidly enough to meet the requirements for many cancer research projects. However, evolving new technologies are helping to make population-based cancer data available in real time. Informatics research has led to refinements in NLP methods. NLP algorithms derive meaning from computerized narrative text documents without additional human intervention.[41,42] A number of population-based cancer surveillance programs are implementing NLP applications in clinical pathology laboratories in an effort to obtain pathology reports for patients with cancer at the time of diagnosis.[43-45] These NLP programs automatically process every pathology report entered into the pathology laboratory computer. If the NLP program determines that the report is associated with a cancer diagnosis, the report is securely transmitted to the registry. Internal audits conducted by several registries have shown that the sensitivity of these systems can approach 100%, and the specificity is commonly around 95%. Implementing NLP programs in all of the clinical pathology laboratories that are covered by a central cancer registry makes population-based rapid case ascertainment a reality.

These systems are critical for studies of patients with cancer who have short survival times. If an investigator wants to conduct a study that requires interviews or the collection of tissue from patients with malignancies such as lung or pancreatic cancer, it will be necessary to have a system in place that facilitates rapid case ascertainment. Automated electronic pathology reporting provides such a system. In addition, if NLP programs have been installed in all of the clinical pathology laboratories that will see histologic material from patients with cancer served by a population-based registry, the sample of cases generated will represent the underlying population, and studies using this system will have strong external validity.

The effort required to install and support NLP reporting from all of the pathology laboratories that will see histologic material from patients with cancer served by a population-based cancer registry is formidable. However, several central cancer registries have already successfully accomplished this, and other population-based cancer registries are in the process of establishing such a system.[15,46]

USING THE POPULATION-BASED CANCER REGISTRY AS A VTR

The ability to perform genomic sequencing on FFPE tissue and the ability to identify the activity levels of many different signaling molecules in FFPE tissue make it possible for the central cancer registry to serve not only as a source for high-quality data but also as a population-based VTR. The central cancer registry has the legal authority to collect all information about each cancer case, including the location of the clinical pathology laboratories where tissue blocks from all cancer patient surgeries and biopsies are stored. As such, it is possible for the registry to serve as an honest broker. The central registry can go through the procedures at each laboratory required to obtain access to the tissue blocks, anonymize the tissue blocks, have the tissue blocks processed in a central research laboratory, return the tissue blocks to the contributing clinical pathology laboratory, and provide investigators with both de-identified high-quality data and population-based tissue samples that lead to studies with strong external validity.

There are a number of examples where central cancer registries have served as VTRs.[47-49] A registry used as a VTR is an especially valuable research resource for basic scientists who are studying specific signaling molecules. Using the central cancer registry as a VTR makes it possible for basic science investigators to determine the activity levels of specific signaling molecules from a sample of cases that truly represent the underlying population and to explore how the expression of these proteins varies by factors such as age, sex, race, treatment, stage, and place. Using the population-based cancer registry as a VTR also makes it possible to explore the activity levels of multiple proteins in different pathways at the same time. This is important because a number of researchers now believe that multiple signaling molecules in different pathways are likely involved in both the onset and recurrence of specific cancers.[50,51]

Using the population-based registry as a VTR requires considerable time and effort. However, because this is such a valuable research resource with the potential to significantly improve cancer research, a number of population-based cancer registries are currently working to develop the infrastructure and establish programs to facilitate using the registry as a VTR.

USING THE POPULATION-BASED CANCER REGISTRY TO GUIDE THE IMPLEMENTATION OF EVIDENCE-BASED CANCER CONTROL INTERVENTIONS AND TO MEASURE CHANGES IN THE CANCER BURDEN AFTER THE IMPLEMENTATION OF THESE INTERVENTIONS

The population-based cancer registry not only is a valuable research resource but also serves as the eyes of our cancer control efforts. The registry makes it possible to see where the cancer incidence rates are high and to measure changes in the incidence rates after the widespread implementation

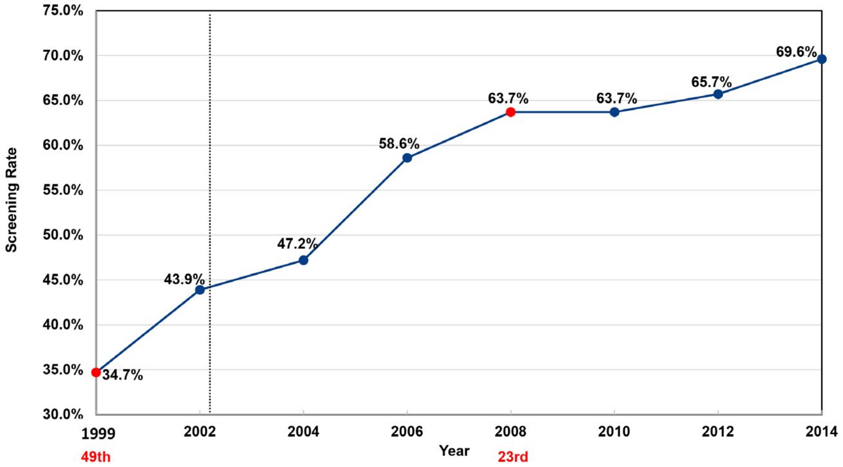


Figure 2. Colorectal cancer screening in Kentucky (1999-2014) from the Behavioral Risk Factor Surveillance System (accessed June 2016). After 1999, the Behavioral Risk Factor Surveillance System colorectal cancer screening questions were asked every other year as of 2002.

of evidence-based interventions that have the potential to reduce the incidence rates. Using the central cancer registry to identify areas and groups with a high cancer incidence and using it to monitor changes in the cancer incidence rates over time have been fundamental surveillance activities of population-based cancer registries for many years. The expansion of high-quality population-based cancer registries now makes it possible for nearly all US states and territories to conduct these critical surveillance activities.

One example of how population-based cancer registry data have been used to identify a high cancer burden and to monitor changes in the incidence rates over time is as follows: Data assembled by the Kentucky Cancer Registry in 2001 identified Kentucky as having the highest colorectal cancer (CRC) incidence rate in the country in comparison with all other US states.[52] In addition, data from the CDC Behavioral Risk Factor Surveillance System showed that Kentucky had the second to lowest CRC screening rate, which was defined as the percentage of the population aged 50 years or older ever having undergone colonoscopy or sigmoidoscopy.[53] These data were used to successfully advocate for a public policy requiring all health insurance companies operating in Kentucky to make screening colonoscopy a covered service for age-eligible individuals.[54] The data were also used by local cancer coalitions to guide the implementation of evidence-based interventions. These interventions included using lay health navigators to reduce cultural barriers and using academic detailing to persuade primary care providers to recommend screening to their age-eligible patients and then schedule CRC screening appointments for these patients.[55-59]

After this intensive effort, the proportion of the age-eligible population in Kentucky undergoing either colonoscopy or sigmoidoscopy rose from approximately one-third (34.7%) in 1999 to nearly two-thirds (63.7%) in 2008, and Kentucky rose from 49th in CRC screening to 23rd.[53] The CRC screening rates in Kentucky have continued to increase.[53] However, as the proportion of the population unscreened increasingly represents hard-to-reach individuals (ie, those who are poor, have lower educational attainment, and have less access to health care), it becomes more difficult to increase the CRC screening rate. The increase in CRC screening in Kentucky from 1999 to 2014 is shown in Figure 2.

Previous studies have shown that increased CRC screening has resulted in a reduced CRC incidence rate through the identification of precancerous polyps and their removal before they become cancer.[60-62] In a large clinical trial conducted by Dr. Wendy Atkin and her colleagues,[60] 112,939 individuals were assigned to the fecal occult blood test, and 57,099 individuals were assigned to flexible sigmoidoscopy. The results of this sentinel study showed that in the intervention group (those undergoing sigmoidoscopy 1 time between the ages of 55 and 64 years), the CRC incidence was reduced by 33%.[60] The purpose of monitoring changes in the CRC burden in Kentucky was to identify variations in CRC incidence rates by age, sex, race, and place and to see whether the widespread implementation of evidence-based interventions designed to increase colorectal screening was accompanied by changes in the CRC incidence rates similar to those found in previous studies.

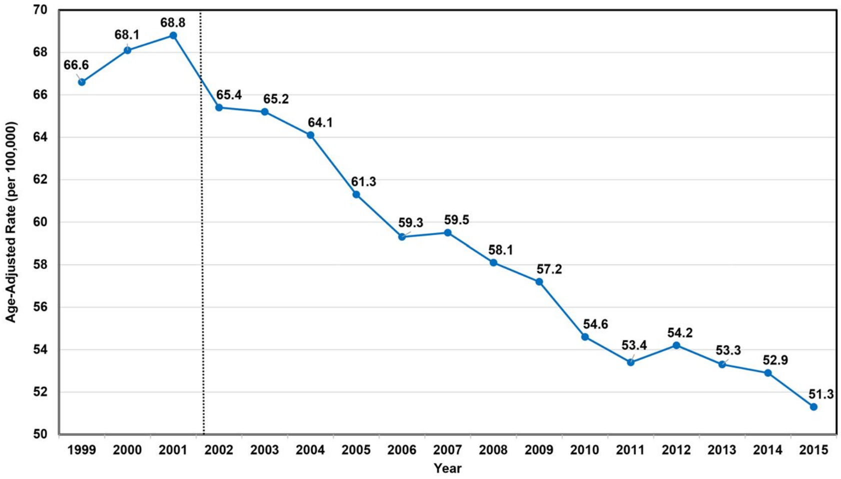


Figure 3. Kentucky colon and rectum cancer incidence rates (1999-2015). All rates are per 100,000 and have been age-adjusted to the 2000 US standard million population.

The increase in CRC screening in Kentucky was accompanied by a 25% decrease in the CRC incidence rate between 2002 and 2015, as shown in Figure 3. Although we cannot say that the increase in CRC screening caused the decrease in CRC incidence, this is consistent with the findings from previous research. The registry was a critical tool both for identifying the problem and for measuring changes in the incidence rates over time. Data from the registry also showed that some geographic areas of the state had a large decrease in CRC incidence, whereas other areas had a much smaller decrease. These findings are essential for helping to direct our limited cancer control resources to areas with the greatest need. Many population-based cancer registries have similar examples of how registry data have been used to identify the burden of specific cancers and to measure changes in the cancer burden over time.[1,63]

DISCUSSION

Modern population-based cancer registries offer a real opportunity to elevate the science of cancer research by providing a sample frame that truly represents the underlying population. Advances in informatics and laboratory science make it possible to obtain complete treatment information, make rapid case identification a reality, and allow the registry to serve as a population-based VTR. Central cancer registries also serve as an essential component of our efforts to measure changes in the cancer burden over time. There are, however, limitations associated with both the technical and scientific advances that help to unlock the potential of population-based cancer registries and with using the registry to measure changes in the cancer burden.

Linking health insurance claims with central cancer registry records provides complete treatment information for each patient, including whether or not the patient received specific genomic tests. However, the results of these genomic tests are not generally available through claims data. Precision medicine increasingly relies on specific genomic signatures to guide the use of appropriate treatment regimens. Not having the results of these tests makes it difficult to determine whether treatment patterns were appropriate from registry data. Some population-based cancer registries are working to collect the results of genomic studies directly from the companies that perform these tests. The NCI SEER cancer registry records are now being linked with the results from Oncotype DX testing performed for patients with breast cancer, and a few central cancer registries are linking their records with companies such as Foundation Medicine, Inc, to obtain the results of tests that examine the presence of a large number of somatic mutations in patients with cancer.[64-66] Information about all of the relevant genomic mutations linked with central cancer registry records would create a remarkable population-based research resource, and it would provide a data set for truly understanding variations in the patterns of care among patients with cancer in the population served by the central cancer registry.

Using the central cancer registry as a VTR provides the framework for conducting population-based studies of tissue from patients with cancer. State laws govern the storage of archival tissue. Nearly all the state laws and clinical pathology laboratories follow the College of American Pathology recommendations. Current College of American Pathology guidelines state that clinical pathology laboratories should retain the FFPE tissue specimens from patient biopsies and surgeries for at least 10 years.[67] Because of limitations on storage space, many pathology laboratories destroy tissue specimens older than 10 years. Therefore, population-based studies using the central cancer registry as a VTR are able to go back only 10 years from the time at which the studies are initiated. Some population-based cancer registries have established discard tissue repositories.[68] These registries receive and store the discarded tissue from clinical pathology laboratories. If a central cancer registry is able to develop agreements to receive the discarded tissue from all of the pathology laboratories that will see histologic material from cancer cases occurring in the geographic area covered by the registry, then the registry can extend the time over which tissue for population-based VTR studies will be available.

The most recent revisions to the Common Rule do not require researchers who are using de-identified archival tissue to obtain informed consent.[69] The body overseeing the revisions to the Common Rule believed that these studies present minimal risk. Furthermore, to put these types of restrictions on the use of archival tissue specimens would greatly limit the ability to use these materials in studies and would likely prevent important discoveries. There are, however, significant ethical considerations. Population-based cancer registries are public health surveillance systems. State laws require that all of the cancer cases diagnosed and/or treated in the geographic area covered by the registry be reported to the registry along with identifying information. Although the rules that govern each population-based registry vary, most registries do not further release patient-identifying information without the patient's permission. The majority of studies that use data from a population-based cancer registry along with archival tissue are void of any patient-identifying information. The use of the data and archival tissue is restricted to studies that have been approved by the registry. In addition, most registries do not provide archival tissue for studies that have a commercial objective.

Central cancer registries make it possible to see population-based changes in the cancer incidence rates after the broad-based implementation of evidence-based interventions. If no substantive changes in the cancer incidence rates are observed, it is clear that the interventions are not having an effect on cancer incidence in the population. However, when significant changes in the incidence rates are observed, it is rarely possible to ascribe these changes to the implementation of the interventions. In other words, it is not possible to say that the interventions caused the observed change in the incidence rates. On the other hand, monitoring changes in the incidence rates over time and using the central cancer registry as a surveillance tool to identify changes in the cancer burden that vary by factors such as age, sex, race, and place are considered essential activities in our efforts to effectively direct cancer prevention and control resources to the areas and groups with the greatest need.

Perhaps the greatest limitation that can diminish our ability to realize the full potential of population-based cancer registries is a lack of funding. Substantial efforts and resources are needed to develop a system for linking all central cancer registry records with all of the public and private administrative health insurance claims, for installing NLP programs in all of the pathology laboratories within the geographic area covered by the registry, and for developing infrastructure that will allow the registry to serve as a VTR. However, central cancer registries offer the most developed and least expensive opportunity to accomplish this. It is far less expensive to make these investments in population-based cancer registries that already capture high-quality, uniform, detailed information on all cancer cases. Despite limited resources, some central cancer registries have already managed to implement these advances.

Understanding how the cancer burden, cancer treatments, and cancer control interventions vary in the population and why these disparities exist is an essential step in our efforts to control the diseases that we classify together as cancer. Central cancer registries offer the opportunity to help accomplish this by providing a set of cases that truly represent the underlying population.

FUNDING SUPPORT

The work described in this article was supported by the National Cancer Institute (Surveillance, Epidemiology, and End Results contract HHSN2612018000131) and the Centers for Disease Control and Prevention (National Program of Cancer Registries cooperative agreement U58DP000810).

CONFLICT OF INTEREST DISCLOSURES

The authors made no disclosures.

REFERENCES

1. White MC, Babcock F, Hayes NS, et al. The history and use of cancer registry data by public health cancer control programs in the United States. *Cancer*. 2017;123(suppl 24):4969-4976.

2. National Cancer Institute. List of SEER registries. https://seer.cancer.gov/registries/list.html. Accessed July 18, 2018.
3. National Cancer Institute. Surveillance, Epidemiology, and End Results Program. https://seer.cancer.gov/. Accessed July 18, 2018.
4. Centers for Disease Control and Prevention. National Program of Central Cancer Registries: about the program. https://www.cdc.gov/cancer/npcr/about.htm. Accessed July 18, 2018.
5. Seiffert J, ed. Standards for Cancer Registries Volume II: Data Standards and Data Dictionary. Sacramento, CA: North American Association of Central Cancer Registries; 1997.
6. Abstracting and Coding Guide for the Hematopoietic Diseases. Rockville, MD: US Department of Health and Human Services; 2002.
7. Ruhl J, Adamo M, Dickie L, Negoita S. Hematopoietic and Lymphoid Neoplasm Coding Manual. Bethesda, MD: National Cancer Institute; 2018.
8. Thronton ML, ed. Standards for Cancer Registries Volume II: Data Standards and Data Dictionary, Record Layout Version 18. Springfield, IL: North American Association of Central Cancer Registries; 2018.
9. Tucker T, Howe H, Weir HK. Certification for population-based cancer registries. *J Regist Manage*. 1999;26:24-27.
10. North American Association of Central Cancer Registries. Certified registries. https://www.naaccr.org/certified-registries/. Accessed November 14, 2018.
11. Bradley CJ, Penberthy L, Devers KJ, Holden DJ. Health services research and data linkages: issues, methods, and directions for the future. *Health Serv Res*. 2010;45:1468-1488.
12. Meray N, Reitsma JB, Ravelli AC, Bonsel GJ. Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. *J Clin Epidemiol*. 2007;60:883-891.
13. Sayers A, Ben-Shlomo Y, Blom AW, Steele F. Probabilistic record linkage. *Int J Epidemiol*. 2016;45:954-964.
14. Carrell DS, Halgrim S, Tran DT, et al. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. *Am J Epidemiol*. 2014;179:749-758.
15. Xie F, Lee J, Munoz-Plaza CE, Hahn EE, Chen W. Application of text information extraction system for real-time cancer case identification in an integrated healthcare organization. *J Pathol Inform*. 2017;8:48.
16. Lewis F, Maughan NJ, Smith V, Hillan K, Quirke P. Unlocking the archive—gene expression in paraffin-embedded tissue. *J Pathol*. 2001;195:66-71.
17. Specht K, Richter T, Muller U, Walch A, Werner M, Hofler H. Quantitative gene expression analysis in microdissected archival formalin-fixed and paraffin-embedded tumor tissue. *Am J Pathol*. 2001;158:419-429.
18. Wu L, Patten N, Yamashiro CT, Chui B. Extraction and amplification of DNA from formalin-fixed, paraffin-embedded tissues. *Appl Immunohistochem Mol Morphol*. 2002;10:269-274.
19. Ahmed EM, Frag AS. Expression of EMMPRIN/CD147 and Ki-67 in oral squamous cell carcinoma: an immunohistochemical study. *J Am Sci*. 2014;10:241-249.
20. Nawal Lutfiyya M, Bhat DK, Gandhi SR, Nguyen C, Weidenbacher-Hoper VL, Lipsky MS. A comparison of quality of care indicators in urban acute care hospitals and rural critical access hospitals in the United States. *Int J Qual Health Care*. 2007;19:141-149.
21. Booth CM, Siemens DR, Li G, et al. Perioperative chemotherapy for muscle-invasive bladder cancer: a population-based outcomes study. *Cancer*. 2014;120:1630-1638.
22. Ma GL, Murphy JD, Martinez ME, Sicklick JK. Epidemiology of gastrointestinal stromal tumors in the era of histology codes: results of a population-based study. *Cancer Epidemiol Biomarkers Prev*. 2015;24:298-302.
23. Cragun D, Weidner A, Lewis C, et al. Racial disparities in BRCA testing and cancer risk management across a population-based sample of young breast cancer survivors. *Cancer*. 2017;123:2497-2505.
24. DeSantis CE, Ma J, Jemal A. Trends in stage at diagnosis for young breast cancer patients in the United States. *Breast Cancer Res Treat*. 2019;173:743-747.
25. DuVall SL, Kerber RA, Thomas A. Extending the Fellegi-Sunter probabilistic record linkage method for approximate field comparators. *J Biomed Inform*. 2010;43:24-30.
26. Noone AM, Lund JL, Mariotto A, et al. Comparison of SEER treatment data with Medicare claims. *Med Care*. 2016;54:e55-e64.
27. Warren JL, Harlan LC, Fahey A, et al. Utility of the SEER-Medicare data to identify chemotherapy use. *Med Care*. 2002;40(8 suppl):IV-55-61.
28. Schrag D, Virnig BA, Warren JL. Linking tumor registry and Medicaid claims to evaluate cancer care delivery. *Health Care Financ Rev*. 2009;30:61-73.
29. Mallin K, Palis BE, Watroba N, et al. Completeness of American cancer registry treatment data: implications for quality of care research. *J Am Coll Surg*. 2013;216:428-437.
30. Hewitt M, Simone JV, eds. Enhancing Data Systems to Improve the Quality of Cancer Care. Washington, DC: National Academies Press; 2000.
31. Hiatt RA, Tai CG, Blayney DW, et al. Leveraging state cancer registries to measure and improve the quality of cancer care: a potential strategy for California and beyond. *J Natl Cancer Inst*. 2015;107:djv047.
32. Foley K, Miller J, Bradley CJ. State-level, population-based cancer registry and administrative data linkages in the United States: a review of the literature [abstract PRM70]. *Value Health*. 2013;16:A25.
33. Boscoe FP, Schrag D, Chen K, Roohan PJ, Schymura MJ. Building capacity to assess cancer care in the Medicaid population in New York State. *Health Serv Res*. 2011;46:805-820.
34. Lipscomb J, Ward KC, Adams K, et al. Augmenting state cancer registry data for quality-of-care assessment: a Georgia-based application to evaluate receipt of adjuvant therapies for breast cancer and colorectal cancer [abstract 5]. *J Clin Oncol*. 2013;31(suppl):5.
35. Kreizenbeck KL, Fedorenko CR, Stickney K, et al. Using cancer registry records linked with health insurance records to measure costs and services at end-of-life [abstract 186]. *J Clin Oncol*. 2016;34:186.
36. National Conference of State Legislators. Collecting health data: all-payer claims databases. http://www.ncsl.org/research/health/collecting-health-data-all-payer-claims-database.aspx. Accessed November 15, 2018.
37. Kehl KL, Lamont EB, McNeil BJ, Bozeman SR, Kelley MJ, Keating NL. Comparing a medical records–based and a claims-based index for measuring comorbidity in patients with lung or colon cancer. *J Geriatr Oncol*. 2015;6:202-210.
38. Hernandez MN, MacKinnon JA, Penberthy L, Bonner J, Huang YX. Enhancing central cancer registry treatment data using physician medical claims: a Florida pilot project. *J Regist Manage*. 2014;41:51-56.
39. Koroukian SM. Linking the Ohio Cancer Incidence Surveillance System with Medicare, Medicaid, and clinical data from home health care and long term care assessment instruments: paving the way for new research endeavors in geriatric oncology. *J Regist Manage*. 2008;35:156-165.
40. Nadpara PA, Madhavan SS. Linking Medicare, Medicaid, and cancer registry data to study the burden of cancers in West Virginia. *Medicare Medicaid Res Rev*. 2012;2:mmrr.002.04.a01.
41. Zeng Z, Shi H, Wu Y, Hong Z. Survey of natural language processing techniques in bioinformatics. *Comput Math Methods Med*. 2015;2015:674296.
42. Hirschberg J, Manning CD. Advances in natural language processing. *Science*. 2015;349:261-266.
43. Penberthy LT, Winn DM, Scott SM. Cancer surveillance informatics. In: Hesse BW, Ahern DK, Beckjord E, eds. Oncology Informatics. Boston, MA: Academic Press; 2016:277-285.
44. Nguyen AN, Moore J, O'Dwyer J, Philpot S. Automated cancer registry notifications: validation of a medical text analytics system for identifying patients with cancer from a state-wide pathology repository. *AMIA Annu Symp Proc*. 2016;2016:964-973.
45. Flagg EW, Datta SD, Saraiya M, et al. Population-based surveillance for cervical cancer precursors in three central cancer registries, United States 2009. *Cancer Causes Control*. 2014;25:571-581.
46. Osborne JD, Wyatt M, Westfall AO, Willig J, Bethard S, Gordon G. Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning. *J Am Med Inform Assoc*. 2016;23:1077-1084.

47. Goodman MT, Saraiya M, Thompson TD, et al. Human papillomavirus genotype and oropharynx cancer survival in the United States of America. *Eur J Cancer*. 2015;51:2759-2767.
48. Saraiya M, Unger ER, Thompson TD, et al. US assessment of HPV types in cancers: implications for current and 9-valent HPV vaccines. *J Natl Cancer Inst*. 2015;107:djv086.
49. Hernandez BY, Goodman MT, Unger ER, et al. Human papillomavirus genotype prevalence in invasive penile cancers from a registry-based United States population. *Front Oncol*. 2014;4:9.
50. Nowell PC. Tumor progression: a brief historical perspective. *Semin Cancer Biol*. 2002;12:261-266.
51. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. *Nat Med*. 2004;10:789-799.
52. Centers for Disease Control and Prevention. United States Cancer Statistics. https://www.cdc.gov/cancer/npcr/uscs/index.htm. Accessed July 12, 2017.
53. Centers for Disease Control and Prevention. Behavioral Risk Factor Surveillance System: prevalence data and analysis tools. https://www.cdc.gov/brfss/data_tools.htm. Accessed July 17, 2017.
54. Kentucky Legislature. Kentucky Revised Statutes Chapter 304.17A-257: coverage under health benefit plan for colorectal cancer examination and laboratory tests. https://apps.legislature.ky.gov/law/statutes/statute.aspx?xml:id=44718. Accessed November 14, 2018.
55. Maxwell AE, Danao LL, Cayetano RT, Crespi CM, Bastani R. Implementation of an evidence-based intervention to promote colorectal cancer screening in community organizations: a cluster randomized trial. *Transl Behav Med*. 2016;6:295-305.
56. Studts CR, Tarasenko YN, Schoenberg NE, Shelton BJ, Hatcher-Keller J, Dignan MB. A community-based randomized trial of a faith-placed intervention to reduce cervical cancer burden in Appalachia. *Prev Med*. 2012;54:408-414.
57. Schoenberg NE, Studts CR, Shelton BJ, et al. A randomized controlled trial of a faith-placed, lay health advisor delivered smoking cessation intervention for rural residents. *Prev Med Rep*. 2016;3:317-323.
58. Dignan M, Shelton B, Slone SA, et al. Effectiveness of a primary care practice intervention for increasing colorectal cancer screening in Appalachian Kentucky. *Prev Med*. 2014;58:70-74.
59. Mader EM, Fox CH, Epling JW, et al. A practice facilitation and academic detailing intervention can improve cancer screening rates in primary care safety net clinics. *J Am Board Fam Med*. 2016;29:533-542.
60. Atkin WS, Edwards R, Kralj-Hans I, et al. Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial. *Lancet*. 2010;375:1624-1633.
61. Holme O, Loberg M, Kalager M, et al. Effect of flexible sigmoidoscopy screening on colorectal cancer incidence and mortality: a randomized clinical trial. *JAMA*. 2014;312:606-615.
62. Schoen RE, Pinsky PF, Weissfeld JL, et al. Colorectal-cancer incidence and mortality with screening flexible sigmoidoscopy. *N Engl J Med*. 2012;366:2345-2357.
63. Bouchardy C, Rapiti E, Benhamou S. Cancer registries can provide evidence-based data to improve quality of care and prevent cancer deaths. *Ecancermedicalscience*. 2014;8:413.
64. Roberts MC, Miller DP, Shak S, Petkov VI. Breast cancer–specific survival in patients with lymph node–positive hormone receptor–positive invasive breast cancer and Oncotype DX recurrence score results in the SEER database. *Breast Cancer Res Treat*. 2017;163:303-310.
65. National Cancer Institute. Oncotype DX database. https://seer.cancer.gov/seerstat/databases/oncotype-dx/index.html. Accessed July 17, 2018.
66. National Cancer Institute. Foundation Medicine. https://gdc.cancer.gov/about-gdc/contributed-genomic-data-cancer-research/foundation-medicine/foundation-medicine. Accessed July 17, 2018.
67. College of American Pathologists. College of American Pathologists (CAP) retention of laboratory records and materials. https://www.ncleg.net/documentsites/committees/PMC-LRC2011/December%205,%202012/College%20of%20American%20Pathologist%20Retention%20Policy.pdf. Accessed July 18, 2018.
68. Goodman MT, Hernandez BY, Hewitt S, et al. Tissues from population-based cancer registries: a novel approach to increasing research potential. *Hum Pathol*. 2005;36:812-820.
69. US Department of Health and Human Services. Federal policy for the protection of human subjects ('Common Rule'). https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html. Accessed November 9, 2018.