MONOGRAPH

New Data Resources, Linkages, and Infrastructure for Cancer Health Economics Research: Main Topics From a Panel Discussion

Stacie B. Dusetzina, PhD,[1] Lindsey Enewold, MPH, PhD,[2]
Danielle Gentile, PhD,[3] Scott D. Ramsey, MD, PhD,[4]
Michael T. Halpern, MD, PhD[2,*]

[1]Department of Health Policy, Vanderbilt Medical Center, Nashville, TN, USA; [2]Healthcare Delivery Research Program, National Cancer Institute, Bethesda, MD, USA; [3]Cardinal Health Specialty Solutions, Dublin, OH, USA; and [4]Fred Hutchinson Cancer Center, Hutchinson Institute for Cancer Outcomes Research, Seattle, WA, USA

*Correspondence to: Michael T. Halpern, MD, PhD, National Cancer Institute, 9609 Medical Center Dr, Room 3E342, Bethesda, MD 20892-9762, USA (e-mail: michael.halpern@nih.gov).

Abstract

Although a broad range of data resources have played a key role in the substantial achievements of cancer health economics research, there are now needs for more comprehensive data that represent a fuller picture of the cancer care experience. In particular, researchers need information that represents more diverse populations; includes more clinical details; and provides greater context on individual- and neighborhood-level factors that can affect cancer prevention, screening, treatment, and survivorship, including measures of financial health or toxicity, health-related social needs, and social determinants of health. This article highlights 3 critical topics for cancer health economics research: the future of the National Cancer Institute's Surveillance, Epidemiology, and End Results-Centers for Medicare & Medicaid Services–linked data resources; use of social media data for cancer outcomes research; and multi-site–linked electronic health record data networks. These 3 topics represent different approaches to enhance data resources, linkages, and infrastructures and are complementary strategies to provide more complete information on activities involved in and factors affecting the cancer control continuum. These and other data resources will assist researchers in examining the complex and nuanced questions now at the forefront of cancer health economics research.

A theme discussed throughout the "Future of Cancer Health Economics Research" virtual conference was the importance of existing data resources. These resources include survey results; medical and pharmacy claims and hospital discharge information; cancer registry findings; electronic health record (EHR) information; and, most importantly, linkages between different types of data resources. Without these data resources, the tremendous achievements in cancer health economics research would not be possible.

However, a parallel theme of the conference was the need for more comprehensive data that represent a fuller picture of the cancer care experience. There has been growing interest in building additional linked data sources that represent more diverse populations (eg, younger individuals with cancer) or that include more detailed clinical information. For example, EHRs can provide information on laboratory and genomic test results, detailed medical history, and orders placed for services intended for patients, including those that were not pursued.

Researchers also need more information on individual- and neighborhood-level factors that can affect individuals' health-care decisions and activities related to cancer prevention, screening, treatment, and survivorship, including health-related social needs and social determinants of health.

In a panel titled "What do we need to be successful? New data infrastructures, resources, and linkages," several researchers discussed the future of data use and data needs for cancer health economics research. Three key topics from this discussion are discussed in the sections below.

The Future of SEER-CMS–Linked Data Resources

In 1991, the National Cancer Institute's (NCI's) Surveillance, Epidemiology, and End Results (SEER) data, which include demographic, clinical, and cause of death information for persons newly diagnosed with cancer as reported to select population-based cancer registers from across the nation, were first linked to Centers for Medicare & Medicaid Services (CMS) data. Over the next 3 decades, 4 SEER-CMS data linkages were created (1). Three of the linkages are Medicare focused: SEER-Medicare, SEER- Medicare Health Outcomes Survey, and SEER-Consumer Assessment of Healthcare Providers and Systems (2,3). Table 1 provides further comparison of the persons and data included in these Medicare-focused linkages. The final linkage is SEER-Medicaid, which includes Medicaid beneficiaries from all 50 states and the District of Columbia who are also in the SEER data (4). The resulting SEER-CMS linkages are essential to understanding cancer care and health outcomes in the United States; analyses of these linked data have resulted in over 2200 publications (5).

Researchers tend to be most familiar with the Medicare-focused SEER-CMS linkages, particularly that fee-for-service (FFS) claims data from hospitals, hospices, home health, durable medical equipment, and individual providers are available starting in 1999 and detailed prescription information for persons enrolled in Part D plans are available starting in 2007. Researchers may be less aware that clinical assessments from the Minimum Data Set (https://resdac.org/cms-data/files/mds-30) and the Outcome and Assessment Information Set (https://resdac.org/cms-data/files/oasis) files for persons enrolled in nursing homes and home health care, respectively, are now requestable through the 3 Medicare-focused SEER-CMS linkages. Additionally, 3 ancillary Part D files were recently made available (https://resdac.org/cms-data/files/pde). The Part D plan characteristics file provides information on plan type, premiums, and cost-sharing. The Part D prescriber characteristics file provides demographics and specialty information on the prescribing provider. Finally, the Part D pharmacy characteristics file provides information on pharmacy type (eg, mail-order, independent, or chain).

Over the next few years, NCI plans to further enhance the SEER-CMS linkages so that more of the cancer patients found in both SEER and CMS data have sufficient information to be included in research analyses. For example, historically, most SEER-Medicare analyses have been restricted to persons enrolled in FFS plans because these are the persons for whom claims data and, therefore, detailed information on comorbidities, treatments, and outcomes have been available. Enrollment in Medicare Advantage plans now represent 42% of the Medicare population, and this percentage is projected to increase to 51% by 2030; the representativeness of analyses that only include FFS enrollees is diminishing (6). Therefore, NCI is

currently evaluating how best to incorporate Medicare Advantage (ie, Medicare Part C) encounter data to provide insights into cancer care and health outcomes among beneficiaries enrolled in managed care. NCI also plans to expand the SEER-Medicaid linkage. Currently, the SEER-Medicaid linkage includes only persons in the SEER data from 2006 to 2013 linked to their Medicaid enrollment data from the same years. Over the next year, NCI will be expanding this linkage to include cancer diagnoses and Medicaid enrollment data to cover 1999-2019. Further, NCI will be evaluating how best to incorporate Medicaid claims data to better understand how these data can provide insights into cancer care and health outcomes among persons enrolled in Medicaid. Based on the initial assessments of the Medicare encounter data and Medicaid claims data, NCI will establish best practice guidelines for using these data before making the data available for request.

NCI also plans to expand the utility of the SEER-CMS linkages for assessing associations between social- and economic-related factors and cancer presentation, treatment, and health outcomes among cancer survivors. Although NCI currently releases many area-level (census tract and Zip code) measures, such as median household income, percent of the population with a high school degree, and racial or ethnic make-up of the population, NCI has plans to include additional area-level measures, such as percent of population who are employed, uninsured, and have transportation and internet access (7). Moreover, because area-level measures can only approximate personal characteristics, NCI is investigating mechanisms that will allow individual-level social- and economic-related measures to be requestable via the SEER-CMS linkages. To include individual-level measures, the SEER-CMS data must be linked to other data resources using personally identifiable information, such as social security number, name, and date of birth. Understandably, sharing of personally identifiable information requires a higher level of scrutiny and coordination than incorporation of area-level measures. Nonetheless, NCI recognizes the importance of conducting research at an individual level and is thus pursuing individually linkable data. For example, NCI and the Department of Housing and Urban Development are creating a linkage between the SEER data, and by extension the SEER-CMS data, and Department of Housing and Urban Development administrative data that will allow assessment of how an individual's receipt of housing assistance affects their cancer care and health outcomes.

Because provider characteristics can affect cancer care and patient health outcomes, NCI is also investigating ways to incorporate more provider-level data within the SEER-CMS linkages. Although provider identifiers (eg, National Provider Identifiers for institutions and individuals, including physicians and nurses) are encrypted in the released data, NCI has access to unencrypted provider identifiers, which allows for data linkages at the provider level. NCI already releases a file with hospital characteristics (eg, bed size and NCI Designated Cancer Center status) and the CMS-created Medicare Data on Provider Practice and Specialty file, which includes information on practice type, provider specialty, and summary Medicare utilization measures (8). NCI also has a mechanism to allow researchers to link to the American Medical Association Physician Masterfile (eg, as another source of physician specialty data) (9). Moving forward, NCI is investigating ways to allow information from the National Plan and Provider Enumeration System (https://npiregistry.cms.hhs.gov/), which is the National Provider Identifier repository and also includes practice and specialty information. The SEER-CMS linkages will soon include

Table 1. Comparison of persons and data included in the 3 Medicare-focused SEER-CMS linkages

	SEER-Medicare	SEER-MHOS	SEER-CAHPS
Persons included			
Medicare fee-for-service enrollees	x		x
Medicare Advantage enrollees	x	x	x
Noncancer comparison group[a]	Medicare 5%	MHOS respondents	CAHPS respondents
Data included			
SEER cancer registry data	x	x	x
Medicare enrollment	x	x	x
Medicare claims data[b]	x		x
Medicare part D prescription drug data	x	x	x
MDS clinical assessment data for nursing home residents	x	x	x
OASIS clinical assessment data for home health enrollees	x	x	x
Physician and hospital characteristics	x		x
CAHPS experience of care survey data			x
MHOS quality of life survey data		x	

[a]Persons not included in the SEER data who resided in the geographic catchment areas for the registries. CAHPS = Consumer Assessment of Healthcare Providers and Systems; CMS = Centers for Medicare and Medicaid Services; MDS = Minimum Data Set; MHOS = Medicare Health Outcomes Survey; OASIS = Outcome and Assessment Information Set; SEER = Surveillance Epidemiology, and End Results.
[b]From in- and out-patient hospitals, hospice agencies, home health agencies, individual providers, and durable medical equipment providers.

information on provider participation in payment and delivery system models (eg, accountable care organizations), and NCI plans to create summary measures based on the SEER-CMS data (eg, for each provider summarize the number and characteristics of patients they serve). Finally, NCI is investigating mechanisms to incorporate other provider characteristics, such as practice size, and participation in health-care networks or compendiums.

NCI is also striving to shorten the time lag between when a new cancer is diagnosed and when the associated SEER-CMS data are made available to researchers. Currently, there is a 2-year period between when a cancer is diagnosed and when the registries submit their data to NCI (eg, cancers diagnosed in 2019 were submitted to NCI in November 2021). There is an additional delay in the release of SEER-CMS data because the linkage process (between SEER and CMS data) occurs every other year and takes approximately 1 year to complete (eg, cancers diagnosed in 2018-2019 will first be released in SEER-CMS by the end of 2022). To release more timely data, NCI is currently investigating the completeness of SEER data submitted 1 year after diagnosis. NCI is also investigating mechanisms to shorten the SEER-CMS linkage process.

Finally, NCI is pursuing mechanisms to make the SEER-CMS data more accessible and easier to analyze. Currently, the data files, particularly the claims files, are large, cumbersome, organized based on place of service or provider (eg, hospital inpatient or outpatient, or hospice), and require extensive programming and coding knowledge to analyze. NCI is therefore developing new SEER-CMS research "products" that will repackage the raw claims data into more analytically friendly data. For example, SEER-Medicare Condensed Resource files will be processed claims data repackaged based on type of care (eg, systemic therapy, radiation, surgery) or type of measure (eg, comorbidity, cost) and will include a reduced set of variables (eg, only dates, diagnosis codes, and procedure codes). Data will then be further simplified to create SEER-Medicare databases that include the SEER data and Medicare time-fixed variables (eg, receipt of systemic therapy within 6 months or 12 months of cancer diagnosis). The resulting SEER-Medicare databases will be accessible through SEER*stat, the statistical software via

which researchers currently access the SEER data. It is anticipated that these new data products along with the underlying coding algorithms will be released within the next 2 years.

In summary, over the years, many enhancements have been made to the SEER-CMS linkages, and more are on the horizon. The new enhancements should expand the generalizability of SEER-CMS data and allow for more timely assessments for a wider range of research topics.

Use of Social Media Data for Cancer Economics Research

Social media data may be a valuable resource for oncology research. Twitter, specifically, has been used by researchers to perform content analysis on a variety of oncology-related topics. Examples include lung cancer prevention and control [10], breast cancer screening [11], engagement with skin cancer prevention public health messaging [12], and social media conversations following celebrity cancer announcements [13]. Researchers have also examined content in Facebook support groups for breast cancer [14], ovarian cancer [15], and caregivers of children with cancer [16]. Struck et al. [17] reviewed multiple social media platforms, including Facebook, Twitter, YouTube, and Instagram, for online prostate cancer communities.

Twitter data are particularly useful for qualitative emergent thematic analysis that can help to illuminate why a certain phenomenon is occurring and the nature of the phenomenon [18,19]. Twitter data are used most frequently in research (compared with other social media data) due in part to their concise messages that can often be summarized into 1 or 2 themes [20]. The short tweets allow for examination of the breadth of a topic with limited depth. Furthermore, tweets are free and publicly available. With the facilitation of software, thousands of tweets can be collected within minutes. The majority of tweets are written by private individuals expressing their opinions [21] and therefore provides insight into lay discourse. In contrast, use of other social media platforms such as Facebook for research may be more complex due to their free-form content, which can be lengthy and contain many themes. The advantages and

limitation of Twitter, Facebook, and other social media platforms as research data sources were previously discussed (20,22).

On Twitter, Tweets are organized with hashtags (#), which group together all messages that include a given hashtag, making it easy to search by topic. For example, if using these data for cancer health economics research, one might search for #LungCancer and #FinancialToxicity or #LungCancer and #Cost. Spending time immersed in tweets on the topic may help researchers uncover hashtags that are common in their subject area. Once the tweets of interest are collected, researchers can interpret the data for emergent themes among the discourse. The data can be reviewed continually until no new themes arise, known as the point of data saturation (23-25). Once saturation is met, the researcher can draw inferences about public opinion on the phenomenon of interest. Insights gained from social media data may also be particularly useful for hypothesis generation in oncology economic research (26). Although using social media data for oncology economic research is nascent and the purpose of this discussion is to highlight its potential for future research, examples of previous research include examination of online crowdfunding campaigns posted on social media, which highlight the financial hardship imposed by kidney cancer (27), and young adult cancer survivors' experiences with economic distress and financial toxicity during the COVID-19 pandemic (28).

Although Twitter data may be useful for research insights, they are not without limitations. Twitter users tend to be younger, more politically democratic (29), and have higher income and education levels compared with the general US population (30). Only data from those accounts visible to the public can be viewed and gathered. Additionally, publicly available Twitter data may not be an accurate or representative picture of the cancer patient experience because the private realities of individuals' experiences that are not publicly shared cannot be accounted for. There may be bias in those whose accounts are public versus private. The validity of the data may also be questionable due to the inability to be certain whether the content is authentic or possibly fabricated or from duplicate accounts. Ultimately, the researcher's interpretation of the content is subject to personal biases.

Social media platforms provide free, publicly available, nonproprietary sources of public discourse that can be mined for emergent qualitative themes on a broad variety of topics and are a novel data source for cancer economics research. Researchers may wish to consider social media data in a formal capacity with emergent thematic qualitative research or informally as a means to draw insights on public opinion. Social media may also be an effective means of disseminating scholarly research findings. A growing number of peer-reviewed journals now ask authors to provide their Twitter username and an example tweet during the submission process that can be used to promote the article on publication.

Multi-Site–Linked EHR Data Networks

Use of multi-site linked EHR and administrative claims data is a recent and promising development for cancer health economics research. Recent, large-scale initiatives to aggregate clinical data across health systems or payers include ASCO's CancerLinQ, Flatiron Health, the HMO Research Network, the National Patient-Centered Clinical Research Network (PCORnet), the Population-based Research to Optimize the Screening Process, Improving the Management of Symptoms During and Following Cancer Treatment, and many others. These aggregated data sources provide important new opportunities to pursue studies of rare conditions or underserved populations for which an individual hospital or health system may not have adequate sample size. Furthermore, recent efforts to create and improve processes for data linkage allow for a more complete view of patient health services use and standardized spending measures from administrative claims. Both can be critical for pursuing cancer health economics and outcomes research projects. Below, we describe challenges and opportunities with using EHR data for cancer health services and health economics research.

Completeness of EHR-Derived Data

One key challenge with using EHRs is that these studies are often limited to a single site of care. For single-institution studies, common critiques are that study findings are not generalizable or may result in too small of a sample for any one cancer type or treatment of interest. Perhaps most important when using electronic health data alone is to recognize the potential for missing data (31). For example, patients may travel quite far to receive initial treatment from an academic medical center or a Comprehensive Cancer Center, and they may receive subsequent health care services outside of the system in which their original care was documented. The experiences of those individuals may not be observed when using only EHR data. This is a particularly high risk for events such as emergency department visits and hospitalizations, which would be more likely to occur close to a patient's home (32). Because of the gaps that exist in each data source individually, there is interest in building multi-site–linked cohorts that can address novel questions regarding patient access to, use of, and the costs associated with cancer care.

Developing EHR Data Standards

For the remainder of this section, we focus our comments on one of the coauthor's recent experiences using the PCORnet multi-site EHR data linked to fee-for-service Medicare data. Recent initiatives to standardize data through the PCORnet have provided cancer health services researchers with a unique opportunity to pursue multi-site studies (33). The PCORnet includes 9 large clinical research networks, representing a diverse set of patients and institutions (34). Each site within the PCORnet follows a set of data standards that allow for similar recording of EHR variables in a consistent manner (35). Researchers may then develop a query at one site that can be distributed to partner sites and efficiently build new cohorts for research studies (36). This provides a unique opportunity to evaluate processes and outcomes of care across diverse sites.

In a recent project, the team led by Vanderbilt University School of Medicine created a first of its kind multi-site–linked cohort from 4 PCORnet clinical research networks, representing 11 different health care systems (37). The intent of this project was to identify individuals who had a prescription order for an anticancer drug (or other high-priced specialty medication) and to verify whether the patient filled that medication within a specified timeframe. The team was particularly interested in estimating the association between expected out-of-pocket costs for the initial fill and patient medication uptake, comparing Medicare beneficiaries with versus without low-income subsidies for medication costs. The network linked the EHR data from participating sites to Medicare FFS claims for all adults

aged 65+ years who received a prescription of interest in each system's EHR.

Barriers to EHR Network Research

There were several challenges encountered in pursuing and developing this network. First, although proposed novel multi-site data linkages were viewed as innovative, questions of feasibility reduced fundability of many such proposals through traditional NIH funding mechanisms (eg, R01s), as experienced by study collaborators. For those primarily interested in studying economic outcomes, funding was also likely constrained over more recent years due to limits on economics-focused research within PCORI (38) and priorities set by the NIH (where priorities focused on applications in which economic outcomes were clearly related to health outcomes rather than considered as primary outcomes themselves) (39). Ultimately, this work was supported by the Commonwealth Fund and the Leukemia & Lymphoma Society, both of which have interests in health-care spending by payers and patients. Additionally, projects proposing to use data from multiple sites faced considerable budget constraints if engaging an investigator and the informatics team at each data-contributing site, which may make them too expensive or lengthy for funding mechanisms specifically geared towards projects with less certain feasibility (eg, R21s). Though these costs are critical to studies that rely on local investigators for manual data collection or other activities, they may be less essential for conducting simple data extractions from the EHR when using the predefined common data model elements. The Vanderbilt team worked with local PCORnet leadership to identify a path forward for this work by first engaging with research partners at sites that were willing to participate in exchange for a fixed fee and for future collaboration on the research products produced. Though this initial effort required substantial unfunded time to develop and establish a process for future studies, the Vanderbilt team is hopeful that it may serve as a roadmap for future studies and as proof of concept for establishing feasibility for R01-funded projects in the future. Alternatively, the NCI may be interested in encouraging data linkage projects like this through other funding mechanisms such as U01 awards, where infrastructure costs may be more readily supported, particularly when such awards would allow for reuse of the data source.

Second, navigating legal and institutional review board processes for the work was an important yet time-consuming process. There are opportunities to streamline legal agreements and institutional review board submissions for data linkage projects in the future. For example, one important concern that made legal agreements more complex was the transfer of private health information. These unique identifiers make easy work of data linkage but must be carefully protected (40). To help derisk these activities, the Vanderbilt team developed a process with the assistance of the CMS and the CMS Research Data Assistance Center for transferring and linking data using a synthetic identifier. This allowed data-contributing sites to send private health information to only a single entity (General Dynamics, the data linkage partner for CMS) rather than send this information to another health system or an individual investigator. This process is now being used in subsequent studies using the PCORnet for data linkage. Ultimately, this process may result in more standardized data use and reuse agreements between sites that encourage greater sharing of linked cohorts.

Discussion

High-quality databases that represent the scope of the cancer patient experience are a critical component of cancer health economics research. Historically, databases that include insurance claims or claims linked to cancer registry records have been the core resources for cancer economics and outcomes researchers. The drive for new data resources that capture a fuller picture of the cancer patient experience reflects the knowledge that insurance claims and registry records address only a subset of the complex and nuanced questions that are now at the forefront of cancer health economics research. To address new questions, a variety of different strategies are being used to access other data resources for cancer populations. Leaders of existing resources are incorporating new data items and databases that collect information on current and novel topics to supplement these important resources.

Linking complimentary data resources permits sharing detailed information from different sources, gaining advantages from both sets of knowledge. The discussions of the *Future of SEER-CMS Linked Data Resources* and the *Multi-Site Linked EMR Data Networks* highlight 2 different approaches for linking data resources, involving 2 very different types of data. As cancer health economics researchers continue to need both greater clinical details and more information on individual and contextual information (eg, neighborhood and regional characteristics, relevant health policies and regulations), linking of disparate data sources will become more critical.

A further strategy to understand the cancer patient experience uses data resources that are not frequently used in economics research. The discussion of *Social Media Data for Cancer Health Economics Research* emphasizes how a ubiquitous data resource can provide rich insights for cancer researchers.

Finally, in some situations, it may be necessary to create a new data resource focused specifically on economic information not available elsewhere. Examples include measures of patient income and assets, patient financial support sources, or practice-level service costs. In addition, few data resources provide information on measures of financial health or toxicity, a key component when examining cancer treatment decision making and the impacts of these decisions. Although development of new data resources occurs frequently in clinical research (eg, every new clinical trial can be considered a new data resource) and to some degree in epidemiology (every new cohort study), there are few new data collection efforts in which the main focus is to gather information for cancer health economics research. This is understandable, because new data collection efforts require substantial effort and cost, and cancer health economics research can generally "piggy-back" off of other data resources. However, it may be worthwhile to consider when new data collection efforts for cancer health economics research are appropriate; for example, when do prospective clinical or observation cancer studies (in any phase of the cancer control continuum) require separate economic components? A discussion of this topic is beyond the scope of this article but illustrates the need for continued focus on data resources, linkages, and infrastructure to further the development of this field.

Notes

References

1. Enewold L, Parsons H, Zhao L, et al. Updated overview of the SEER-Medicare data: enhanced content and applications. *J Natl Cancer Inst Monogr*. 2020; 2020(55):3–13.
2. Ambs A, Warren JL, Bellizzi KM, et al. Overview of the SEER–Medicare health outcomes survey linked dataset. *Health Care Financ Rev*. 2008;29(4):5–21.
3. Chawla N, Urato M, Ambs A, et al. Unveiling SEER-CAHPS(R): a new data resource for quality of care research. *J Gen Intern Med*. 2015;30(5):641–650.
4. Warren JL, Benner S, Stevens J, et al. Development and evaluation of a process to link cancer patients in the SEER registries to national Medicaid enrollment data. *J Natl Cancer Inst Monogr*. 2020;2020(55):89–95.
5. National Cancer Institute. Division of Cancer Control and Population Sciences, Program HDR. Search SEER-CMS linkages publications. https://healthcaredelivery.cancer.gov/publications/. Accessed November 29, 2021.
6. Freed M, Fuglesten Biniek J, Damico A, et al. Medicare Advantage in 2021: enrollment update and key trends. https://www.kff.org/medicare/issue-brief/medicare-advantage-in-2021-enrollment-update-and-key-trends/. Accessed June 21, 2021.
7. National Cancer Institute, Division of Cancer Control and Population Sciences, Program HDR. SEER-Medicare: geographic area data. https://healthcaredelivery.cancer.gov/seermedicare/aboutdata/geographic.html. Accessed May 3, 2022.
8. (ResDAC) RDAC. Medicare data on provider practice and specialty (MD-PPAS). https://resdac.org/cms-data/files/md-ppas. Accessed May 3, 2022.
9. American Medical Association. AMA physician masterfile. https://www.ama-assn.org/about/masterfile/ama-physician-masterfile. Accessed May 3, 2022.
10. Sutton J, Vos SC, Olson MK, et al. Lung cancer messages on Twitter: content analysis and evaluation. *J Am Coll Radiol*. 2018;15(1 Pt B):210–217.
11. Nastasi A, Bryant T, Canner JK, et al. Breast cancer screening and social media: a content analysis of evidence use and guideline opinions on Twitter. *J Cancer Educ*. 2018;33(3):695–702.
12. Nguyen J, Gilbert L, Priede L, et al. The reach of the "Don't Fry Day" Twitter campaign: content analysis. *JMIR Dermatol*. 2019;2(1):e14137.
13. Vos SC, Sutton J, Gibson CB, et al. Celebrity cancer on Twitter: mapping a novel opportunity for cancer prevention. *Cancer Control*. 2019;26(1): 1073274819825826.
14. Bender JL, Jimenez-Marroquin MC, Jadad AR. Seeking support on Facebook: a content analysis of breast cancer groups. *J Med Internet Res*. 2011;13(1):e16.
15. Erfani SS, Abedin B, Daneshgar F. A qualitative evaluation of communication in ovarian cancer Facebook communities. Presented in: *International Conference on Information Society (i-Society 2013)*. June 24-26, 2013; Toronto, Canada.
16. Gage-Bouchard EA, LaValley S, Mollica M, et al. Cancer communication on social media: examining how cancer caregivers use Facebook for cancer-related communication. *Cancer Nurs*. 2017;40(4):332–338.
17. Struck JP, Siegel F, Kramer MW, et al. Substantial utilization of Facebook, Twitter, YouTube, and Instagram in the prostate cancer community. *World J Urol*. 2018;36(8):1241–1246.
18. Elo S, Kyngäs H. The qualitative content analysis process. *J Adv Nurs*. 2008; 62(1):107–115.
19. Krippendorff K. *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: SAGE; 2013.
20. Blagus N, Žitnik S. Social media comparison and analysis: the best data source for research? In: *2018 12th International Conference on Research Challenges in Information Science (RCIS)*, May 29–31, 2018: 1–10; Nantes, France. doi:10.1109/RCIS.2018.8406662.
21. Hays R, Daker-White G. The care.data consensus? A qualitative analysis of opinions expressed on Twitter. *BMC Public Health*. 2015;15:838–838.
22. Giglietto F, Rossi L, Bennato D. The open laboratory: limits and possibilities of using Facebook, Twitter, and YouTube as a research data source. *J Technol Hum Serv*. 2012;30(3-4):145–159.
23. Bowen GA. Naturalistic inquiry and the saturation concept: a research note. *Qual Res*. 2008;8(1):137–152.
24. Guest G, Bunce A, Johnson L. How many interviews are enough? An experiment with data saturation and variability. *Field Methods*. 2006;18(1):59–82.
25. Morse JM. The significance of saturation. *Qual Health Res* . 1995;5(2):147–149.
26. D'Souza RS, Hooten WM, Murad MH. A proposed approach for conducting studies that use data from social media platforms. *Mayo Clin Proc*. 2021;96(8): 2218–2229.
27. Thomas HS, Lee AW, Nabavizadeh B, et al. Characterizing online crowdfunding campaigns for patients with kidney cancer. *Cancer Med*. 2021;10(13): 4564–4574.
28. Thom B, Benedict C, Friedman DN, et al. Economic distress, financial toxicity, and medical cost-coping in young adult cancer survivors during the COVID-19 pandemic: findings from an online sample. *Cancer*. 2021;127(23):4481–4491.
29. Pew Research Center. Differences in how Democrats and Republicans behave on Twitter. https://www.pewresearch.org/politics/2020/10/15/differences-in-how-democrats-and-republicans-behave-on-twitter/. Accessed May 3, 2022.
30. Pew Research Center. Sizing up Twitter users. https://www.pewinternet.org/wp-content/uploads/sites/9/2019/04/twitter_opinions_4_18_final_clean.pdf. Accessed May 3, 2022.
31. Haneuse S, Arterburn D, Daniels MJ. Assessing missing data assumptions in EHR-based studies: a complex and underappreciated task. *JAMA Netw Open*. 2021;4(2):e210184.
32. Patient-Centered Outcomes Research Institute. Data quality and missing data in patient centered outcomes research using emr/claims data meeting summary. https://www.pcori.org/sites/default/files/PCORI-Data-Quality-and-Missing-Data-Workgroup-Summary-121015.pdf. Accessed May 3, 2022.
33. Curtis LH, Brown J, Platt R. Four health data networks illustrate the potential for a shared national multipurpose big-data network. *Health Aff (Millwood)*. 2014;33(7):1178–1186.
34. PCORnet Clinical Research Networks. https://pcornet.org/network/. Accessed May 3, 2022.
35. Rosenbloom ST, Carroll RJ, Warner JL, et al. Representing knowledge consistently across health systems. *Yearb Med Inform*. 2017;26(1):139–147.
36. Corley DA, Feigelson HS, Lieu TA, et al. Building data infrastructure to evaluate and improve quality: PCORnet. *J Oncol Pract*. 2015;11(3):204–206.
37. Dusetzina SB, Huskamp HA, Rothman RL, et al. Many Medicare beneficiaries do not fill high-price specialty drug prescriptions. *Health Aff (Millwood)*. 2022; 41(4):487–496.
38. Padula WV, McQueen RB. Expanding the role of the Patient-Centered Outcomes Research Institute: reauthorization and facilitating value assessments. *Appl Health Econ Health Policy*. 2019;17(6):757–759.
39. National Institutes of Health. Clarifying NIH priorities for health economics research. https://grants.nih.gov/grants/guide/notice-files/not-od-16-025.html. Accessed July 9, 2021.
40. Dusetzina SB, Tyree S, Meyer AM, Meyer A, Green L, Carpenter WR. *Linking Data for Health Services Research: A Framework and Instructional Guide [Internet]*. Rockville, MD: Agency for Healthcare Research and Quality (US); 2014. Report No.: 14-EHC033-EF. https://www.ncbi.nlm.nih.gov/books/NBK253313/. Accessed May 3, 2022.