Seminars in Oncology Nursing 39 (2023) 151428

Contents lists available at ScienceDirect

Seminars in Oncology Nursing

journal homepage: https://www.journals.elsevier.com/seminars-in-oncology-nursing

# Big Data in Oncology Nursing Research: State of the Science

Carolyn S. Harris[a,†], Rachel A. Pozzar[b,†], Yvette Conley[c], Manuela Eicher[d], Marilyn J. Hammer[e], Kord M. Kober[f], Christine Miaskowski[g], Sara Colomer-Lahiguera[h,*]

[a] Postdoctoral Scholar, School of Nursing, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
[b] Nurse Scientist at Phyllis F. Cantor Center for Research in Nursing and Patient Care Services, Dana-Farber Cancer Institute, Boston, Massachusetts, USA and Instructor at Harvard Medical School, Boston, Massachusetts, USA
[c] Professor, School of Nursing, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
[d] Associate Professor and Director of the Institute of Higher Education and Research in Healthcare (IUFRS), Faculty of Biology and Medicine, University of Lausanne, and Lausanne University Hospital, Lausanne, Switzerland
[e] Director, The Phyllis F. Cantor Center for Research in Nursing and Patient Care Services, Dana-Farber Cancer Institute, Boston, Massachusetts, USA and Lecturer at Harvard Medical School, Boston, Massachusetts, USA
[f] Associate Professor, School of Nursing, University of California, San Francisco, California, USA
[g] Professor, Schools of Medicine and Nursing, University of California, San Francisco, California, USA
[h] Senior Nurse Scientist and Junior Lecturer, Institute of Higher Education and Research in Healthcare (IUFRS), Faculty of Biology and Medicine, University of Lausanne, and Lausanne University Hospital, Lausanne, Switzerland

ARTICLE INFO

*Key Words:*
Big data
Data science
Malignant neoplasms
Nursing research
Oncology nursing

ABSTRACT

*Objective:* To review the state of oncology nursing science as it pertains to big data. The authors aim to define and characterize big data, describe key considerations for accessing and analyzing big data, provide examples of analyses of big data in oncology nursing science, and highlight ethical considerations related to the collection and analysis of big data.
*Data Sources:* Peer-reviewed articles published by investigators specializing in oncology, nursing, and related disciplines.
*Conclusion:* Big data is defined as data that are high in volume, velocity, and variety. To date, oncology nurse scientists have used big data to predict patient outcomes from clinician notes, identify distinct symptom phenotypes, and identify predictors of chemotherapy toxicity, among other applications. Although the emergence of big data and advances in computational methods provide new and exciting opportunities to advance oncology nursing science, several challenges are associated with accessing and using big data. Data security, research participant privacy, and the underrepresentation of minoritized individuals in big data are important concerns.
*Implications for Nursing Practice:* With their unique focus on the interplay between the whole person, the environment, and health, nurses bring an indispensable perspective to the interpretation and application of big data research findings. Given the increasing ubiquity of passive data collection, all nurses should be taught the definition, characteristics, applications, and limitations of big data. Nurses who are trained in big data and advanced computational methods will be poised to contribute to guidelines and policies that preserve the rights of human research participants.
© 2023 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

## Introduction

With advances in technology, the conceptualization, definition, and use of big data in research have evolved. An early definition of big data included three main attributes, known as the three Vs: "high volume, high velocity, and/or high variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."[1] Volume refers to a large amount of data; velocity refers to a high-frequency stream of incoming data; and variety refers to a wide range of data sources or types that require different syntactic formats. Additional Vs that were added over time include variability (ie, the extent to which investigators must differentiate "noise" from important data), veracity (ie, data quality or accuracy), and value (ie, the importance of the data).[2-4]

* Address correspondence to: Sara Colomer-Lahiguera, Institute of Higher Education and Research in Healthcare (IUFRS), Office 01/169 - PROLINE - Rte de la Corniche 10 – 1010 Lausanne, Switzerland.
*E-mail address:* sara.colomer-lahiguera@chuv.ch (S. Colomer-Lahiguera).
† Co-first authors.

https://doi.org/10.1016/j.soncn.2023.151428
0749-2081/© 2023 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

Big data in health care encompasses large amounts and diverse types of data from the rapid and increased digitization of individual patient information. The use of big data to improve health outcomes requires cost-effective collection of information from different sources, conversion and storage of data into specific formats, and processing and analyses of this information according to the needs of the user.[5] Data can be obtained from internal or external sources, including clinical and biological data from electronic health records (EHRs) or research (eg, omics), public or government records (eg, public datasets), or financial records (eg, insurance or payor).[6] In addition, big data includes patient-generated health data (PGHD). PGHD are "health-related data – including health history, symptoms, biometric data, treatment history, lifestyle choices, and other information – created, recorded, or gathered by or from patients (or family members or other caregivers) to help address a health concern."[7] Social media can be a complementary source of health-related data and may be used for epidemiological surveillance or control.[8]

The use of high-volume datasets in nursing research is well established.[9] For decades, nurse scientists have led analyses of data collected as part of routine health care and administration. Several landmark nursing studies have leveraged clinical and administrative claims data to inform safe staffing ratios[10] and approaches to pressure ulcer[11] and fall risk assessment.[12] Leveraging routinely collected data offers an alternative to the collection of large quantities of data directly from research participants, which may impose a burden on some individuals with a health impairment.[13] When large datasets include nurse-sensitive indicators (eg, patient falls, nosocomial infection rates), analyses may provide evidence for the value of nursing care and its association with health outcomes.[14]

Over the years, advances in computing power and computational methods (see Papachristou et al in this Big Data Special Issue) have expanded the potential for high-velocity and high-variety data to meaningfully inform patient care.[15] Oncology nurses tailor their interventions to account for the biological, social, cultural, and environmental factors that may affect a person's well-being. High-variety data have the potential to inform this patient-centered approach. For example, big data often underlies precision health initiatives that aim to deliver health care that is optimized for a person's unique genetic or genomic composition, lifestyle influences, and the context in which they live.[16] Large datasets composed of information from a variety of sources can help oncology nurse scientists identify novel biological, psychosocial, or environmental factors that predict or contribute to disease burden. In addition, analyses of big data may support clinical decision making by identifying complex combinations of factors that predict adverse health outcomes. In turn, these analyses may allow nurses to identify patients who may benefit from proactive interventions.[13] The authors aim to describe some of the most common sources of big data available to oncology nurse researchers, describe access considerations to these data sources; and provide exemplars of big data research from oncology nurse scientists. In addition, the authors describe important ethical issues that need to be considered when amassing, using, and reporting findings from big data analyses and suggest directions for future research.

## Sources of Big Data and Access Considerations

### *Electronic Health Record*

The EHR exemplifies big data. It consists of a large volume of clinically relevant information that is continually updated and derived from a variety of sources. Data stored in the EHR are varied and may include clinician notes, vital signs, laboratory reports, telemetry data, imaging data, ICD codes, and PGHD (eg, symptom reports). Investigators can extract structured data from the EHR to characterize study participants. Structured data have a standardized format and are easily stored in an organized database. Examples of structured data that are relevant to oncology nursing research include date of cancer diagnosis, blood pressure, and tumor stage. Conversely, unstructured data lack a standardized format and are more difficult to organize. Examples of unstructured data include clinician's narratives, scanned handwritten notes or test results, and free-text findings from imaging studies. Because manual review and extraction of unstructured data are time-consuming and costly,[17] these data are currently underused in research. The underrepresentation of unstructured data in the oncology literature represents a missed opportunity, given that an estimated 70% to 80% of EHR data are unstructured.[18]

Novel computational methods have the potential to analyze large volumes of unstructured EHR data efficiently and accurately. For example, in patients with multiple chronic conditions, natural language processing (NLP) was used to analyze and extract symptom data from nursing notes to identify groups of patients with similar symptom cluster profiles.[19,20] In the oncology setting, NLP was used to analyze narrative EHR data from 808 patients receiving palliative care at the end of life.[17] The investigators sought to develop and evaluate models to detect social distress, spiritual pain, and severe symptoms from 1,554,736 clinician narratives. The investigators developed core search terms for each construct, trained NLP models by manually annotating the presence or absence of each construct in a subset of the data, and evaluated each model's performance with the remaining data. Although the NLP models for detecting social distress, spiritual pain, severe pain, dyspnea, and nausea demonstrated high accuracy, those for detecting severe insomnia and anxiety demonstrated moderate accuracy. Although the investigators found that the positive predictive values of the NLP models for detecting social distress and spiritual pain were poor, this finding may reflect the quality of the data recorded. One adage that applies to big data analyses is "garbage in, garbage out," which refers to the importance of training computational models on high-quality data. Nevertheless, NLP is approximately 10 times faster than manual coding and may identify information that human analysts overlook.[21] The development and refinement of additional computational methods in coordination with efforts to promote standardization in clinical documentation will facilitate oncology nurse scientists' ability to leverage unstructured EHR data.

### *Patient-Generated Health Data and Remote Monitoring*

Technological advancements have enabled more powerful and portable personal electronic devices that consumers can wear and/or interact with, producing vast amounts of data. Smartphones, mobile health applications (apps), and wearable devices have increased the frequency, amount, and types of PGHD available. In contrast to clinical data, PGHD allow patients to be responsible for capturing, recording, and deciding whether and with whom to share their data.[7]

PGHD allows a continuous tracing of consumer-specific entries, such as those related to location, physical activity, heart rate, blood pressure, glucose, temperature, sleep patterns, or adherence to medication, among others. Remote longitudinal and real-time monitoring can standardize the collection of data across patients and clinics and decrease information gaps (eg, recent changes in a patient's condition; symptoms that prompt a change in the care plan).[22] In addition, remote digital methods may facilitate retention of and access to a wider and more diverse group of participants, reducing costs and time to create targeted cohort groups, in comparison to traditional clinical studies.[23,24] Furthermore, PGHD may offer cost-effective strategies by optimizing cancer care outside of the clinic.[25] In clinical research, detailed information about the time of collection, amount, or combination of data sources can help to standardize and capture more precise and frequent data to understand mechanisms and toxicities of cancer treatments and improve the efficiency of oncology clinical trials.[26,27] Moreover, predictive models of disease states can be tested and health-promotion interventions created. The use of PGHD enables a shift from provider-driven to patient-led activities

that enables self-monitoring and self-management and fosters patient engagement.[28] However, additional research is needed on the legal, ethical, feasibility, and modeling issues related to the acquisition and use of PGHD.

*Wearable Health Devices (passive reporting)*

A wearable is a device with a sensor that can collect health-related data remotely with the advantage of minimizing discomfort and interference with normal human activities. This approach makes it possible to monitor patients in their own environment.[29] Wearable and remote patient monitoring devices may be fastened to the wrist, upper arm, waist, hip, or other body parts. These devices can provide biometric data, including heart rate, electrocardiogram, respiratory rate, blood oxygen saturation, blood glucose, sleep pattern, and body temperature. The collection of data from wearable and remote patient-monitoring devices can take place in real time or during scheduled data transfers. In this sense, these devices combine the three main Vs of big data: large amounts of data (volume) that are collected in real time or at high frequency (velocity) from a wide range of data sources (variety).

In the oncology setting, several examples exist of the use of wearable and remote patient-monitoring devices to improve patient outcomes during and after cancer treatment. As part of a European project titled Integrated Network for Completely Assisted Senior Citizens' Autonomy (inCASA),[30] a home-based platform was used to monitor real-time symptoms in patients receiving chronomodulated chemotherapy at home.[31] Circadian rest-activity rhythm and sleep were measured with a wrist accelerometer, body weight changes with a dedicated scale, and symptom information with a questionnaire completed on an interactive electronic screen. Evidence for the acceptability of this approach included 5,891 data points collected over 364 patient-days out of the 8,736 expected (67.4%), with a median daily adherence of 73%. This approach allowed a day-to-day multidimensional and accurate evaluation of each patient's response to the treatment and helped document the safety of chronomodulated triplet chemotherapy delivery in the patient's home.

In contrast, other studies reported suboptimal adherence to wearable health devices. The OncoWatch 1.0 study investigated the feasibility of using smartwatches to monitor heart rate and physical activity in patients with head and neck cancer who were receiving radiotherapy.[32] Only 31% of patients adhered to the study protocol that entailed wearing a smartwatch for 12 hours per day during and for 2 weeks after radiotherapy. The investigators proposed that the task of charging the watch and not being able to use the watch for personal purposes led to low adherence.

Another example of the use of sensors for home-based cancer symptom management is Behavioral and Environmental Sensing and Intervention for Cancer (BESI-C).[33] In this study, dyads of patients with cancer and their primary caregivers were followed to monitor cancer pain and distress at home. Environmental sensors assessed the home context (eg, light and temperature), and Bluetooth beacons located dyad positions. Both patients and caregivers wore smartwatches to record and characterize pain events. This study introduced a new approach to monitoring and mitigating the escalation of cancer pain and distress by controlling environmental and contextual factors at home. Participants reported that the intervention was meaningful and not burdensome.

*Patient-reported data (active reporting)*

Patient-reported outcomes (PROs) are systematic ways of measuring patients' subjective views about the impact of their disease and its treatment. From a value-based care point of view, collecting PRO data could help to evaluate, monitor, and improve provider and setting performance, or establish standards and benchmarks to measure the effectiveness of a health system.[34] One study[35] identified three potential uses of "Big PRO" data: (1) to guide individual care through real-time monitoring; (2) to develop population-level prognostic models to predict patients most likely to benefit from an intervention and to identify those who are a priority for care; and (3) to enrich observational research in real-world trials. Despite their established use in clinical trials, PROs are not universally collected in real-world clinical settings. One barrier to the integration of PROs into routine care is that many EHRs are not designed to meaningfully display and assist clinicians to interpret PRO data.[36,37] For nursing, the lack of PROs in EHRs limits the extent to which nursing interventions such as patient education, symptom evaluation, and symptom management can be measured and evaluated.[38]

In 2013, the Patient-Centered Outcomes Research Institute (PCORI) in the United States launched PCORnet, the National Patient-Centered Clinical Research Network, a major initiative to create an effective and sustainable infrastructure to support researchers in learning from clinical and patient-reported outcomes in large observational studies.[39] Another example is the Dutch population-based Patient-Reported Outcomes Following Initial treatment and Long-Term Evaluation of Survivorship (PROFILES) registry, which combines longitudinal PRO measures, objective measures, and cancer registry, ambulatory, and pharmacy data.[40]

In France, the CANcer TOxicities (CANTO) longitudinal cohort study (NCT01993498) is developing a database of chronic treatment-related toxicities in 14,750 women with stage I to III breast cancer.[41] The aims of the study are to quantify the impact of treatment toxicities and to generate predictors of chronic toxicity in patients with nonmetastatic breast cancer. CANTO collects PROs (ie, quality of life, psychological, behavioral), as well as clinical, treatment, toxicity, socioeconomic, and biologic data. These initiatives will allow the full integration of PROs and information related to their impact into EHRs, claims databases, and other sources of big health data. In addition, initiatives such those undertaken by the Organization for Economic Cooperation and Development[42] and the International Consortium for Health Outcomes Measurement[43] aim to support and develop a coherent and comprehensive approach to standardizing and implementing the systematic collection of PRO data internationally.

*Large Public Datasets*

A major challenge faced by researchers is the acquisition of high-quality data. Prospective data collection can be an expensive process that is time intensive for both researchers and patients. Due to funding constraints, researchers must make difficult decisions about what types of data to collect, number of assessments, and number of patients. Furthermore, multiple years pass between the grant writing process and the beginning of data analysis, which impedes progress in oncology research. The availability of publicly available datasets with large samples (eg, >1000 participants) that acquire data longitudinally and include various types of data (eg, symptom severity, gene expression) can accelerate oncology research. To improve the management of data produced by studies funded by the National Institutes of Health (NIH) in the United States and increase the responsible sharing of these data, the Policy for Data Management and Sharing[44] was enacted requiring that researchers of NIH-funded studies share their data with a quality data repository (eg, Database of Genotypes and Phenotypes [dbGaP]). Given that this policy went into effect as of January 2023, data within publicly accessible data repositories will expand in volume exponentially. In addition to the databases previously described (ie, PROFILES, CANTO), the next section of this paper describes five publicly available datasets with a high variety, velocity, and volume of data that oncology nurse scientists can access to explore a variety of research questions.

*National biobanks*

The United Kingdom (UK) Biobank is a biomedical database composed of growing volumes of a variety of data used to identify the

underlying causes (eg, environmental, genetic) of various diseases. Recruitment of more than 500,000 UK citizens took place between 2006 and 2010, and the study continues to prospectively collect data on all living participants.[45] The target age for recruitment was 40 to 69 years because this age period is associated with increased development of various conditions, including cancer. Participants were required to be registered with the universal health care system of the UK, provide consent for long-term follow-up, and allow for study access to their health records. Therefore, detailed health records on cancer and death registry data and inpatient and primary care records are updated annually and are available on all participants.[46] Data available for analyses include detailed questionnaires on health, lifestyle, and exposures; physical measures and accelerometer data; whole genome and exome sequencing on all participants; blood, urine, and saliva for proteomic, metabolomic, and telomere analyses; and magnetic resonance imaging of the brain, heart, and full body. Access to this rich resource is available to the international scientific community through application.

The NIH launched the All of Us Research Program in 2015, recognizing that a "one size fits all" policy for disease prevention and treatment may not be effective for every person.[47] All of Us proposes that to determine the specific risk factors for various diseases and to develop individualized treatments, the influence of one's environment, lifestyle, family history, and genetic makeup on disease development and treatment efficacy must be evaluated. Acknowledging the historic absence and exclusion of people from racial and ethnic minority communities, rural communities, and lower socioeconomic status in biomedical research,[48] All of Us is committed to the recruitment of participants who reflect the diversity of the United States. To date, All of Us is more than halfway to its goal of recruiting 1 million participants and plans to collect additional data over time. Types of data being collected include patient-reported surveys on one's environment, lifestyle, and other social determinants of health; EHR data; physical measures; blood, urine, and saliva samples; and digital health data. While recruitment and data collection are ongoing, current deidentified data can be accessed on three tiers: Public Tier (ie, view data snapshots, no registration required), Registered Tier (ie, includes data from EHRs and surveys, registration required), and Controlled Tier (ie, genomic data, registration and prior approval required).

Using cross-sectional data from 14,127 participants in the All of Us Research Program, symptom phenotypes in participants diagnosed with one or more chronic conditions (ie, cancer, chronic obstructive pulmonary disease, heart failure, and/or type 2 diabetes mellitus) and risk factors that predicted membership in these symptom phenotype groups were evaluated.[49] Cohort Builder within the All of Us Researcher Workbench was used to identify study participants for analysis. Eligible participants were required to have one or more of the prespecified chronic conditions and complete response data on fatigue, emotional distress, and pain items on the Overall Health Survey that was collected after diagnosis. Using hierarchical cluster analysis, four distinct symptom phenotypes were identified (ie, mild symptoms, severe emotional distress, severe pain, severe symptoms). Participants who forwent or delayed medical care or rated their mental or physical health as poor were more likely to belong to the severe emotional distress, pain, or symptom phenotypes.

### *National survey data*

Another type of large, publicly available data that researchers can use is data compiled from national or international surveys. For example, in an effort to improve patient-centered care, health care systems and governments are increasingly using large-scale, population-wide, patient-reported surveys to examine patients' experiences across the cancer-care continuum. These surveys provide a perspective on the aspects of cancer care that patients find most important. Notably, patient-reported experiences complement data on health outcomes (eg, treatment effectiveness, mortality), which together provide a more holistic picture of the quality of health care.[50]

In the United States, the Consumer Assessment of Healthcare Providers and Systems Cancer Care Survey examines patient experiences in the context of their interactions with various clinicians and staff (eg, communication, perceived respectfulness of staff), experiences with health care facilities (eg, care coordination, timeliness of appointments), and perception of overall cancer care.[51] Using a similar survey, the UK uses the Cancer Patient Experience Survey to assess changes in cancer care and as a tool to inform quality improvement.[52] The Patient-Reported Indicator Survey (PaRIS) of People Living with Chronic Conditions measures both patient-reported experiences (eg, care coordination, wait times) and PROs (eg, quality of life, physical functioning) in adults living with one or more chronic conditions.[42] Because PaRIS is an international survey, researchers and institutions can compare data within and across countries. Researchers have used data generated from large-scale patient-experience surveys to examine factors associated with patient care experiences in older patients with hematologic malignancies,[53] associations between having a better care experience with a clinical nurse specialist and overall survival in patients with heterogeneous types of cancer,[54] and variations in patient experiences with cancer care by type of cancer in patients with heterogeneous types of cancer.[55]

### *Social Media*

Worldwide, an estimated 4.74 billion people use social media.[56] Social media platforms allow users to engage with each other and share user-generated content.[57] The most widely used social media platforms include YouTube, Facebook, and Twitter.[58] Social media may be used by individuals to exchange social, emotional, and practical support related to a health condition or to find and share health information.[59] To date, investigators have used social media to recruit research participants rather than as a source of research data. However, investigators may face several challenges related to participant misrepresentation when they use social media platforms for recruitment.[60] Investigators who analyze content that social media users share publicly may avoid these challenges. Although user-generated social media content may shed light on the experiences of people with cancer and other conditions, the unstructured nature of this content has limited the extent to which it has been formally analyzed.

Online discussion forums represent an especially promising source of high-velocity unstructured health data. In a study that aimed to develop an automated model to classify the needs expressed by patients and caregivers online,[61] 853 messages shared in an online health community for people with ovarian cancer and their caregivers were analyzed. First, messages that referenced physical, psychological, social, and information needs were manually annotated. Next, a machine learning model that used a "bag of words" representation was built, using the combination and frequency of the words in each message to predict the needs expressed in each message. The resultant classification model was able to identify different types of needs with a high level of accuracy. These findings suggest that novel computational methods such as machine learning are a feasible approach to use to analyze large amounts of unstructured user-generated data.

### *Omics*

To determine the complex mechanisms that underlie common symptoms in patients with cancer, oncology nurse scientists are increasingly incorporating omics approaches to their research. The various types of omics data can be conceptualized as levels of biological data (eg, genomics, transcriptomics, proteomics). Given that each type of omics data provides valuable and unique insights into the molecular underpinnings of various conditions, researchers may

select one or more types of omics data for their analyses based on their research questions and/or hypotheses.[62] For example, epigenomics data (eg, DNA methylation) can be used to examine linkages between social determinants of health and symptom or health outcomes.[63] Findings from these studies have the potential to identify biomarkers of disease or symptoms and lead to the development of tailored and targeted interventions.

For example, an interdisciplinary team of oncology nurse and physician researchers, bioinformaticians, and molecular geneticists integrated a variety of high-volume data types to identify a potential target for intervention in breast cancer survivors with paclitaxel-induced peripheral neuropathy. In their first study,[64] a transcriptome-wide differential gene expression analysis (11,487 genes) was performed between breast cancer survivors who did (n=25) and did not (n=25) develop paclitaxel-induced peripheral neuropathy as a result of paclitaxel administration. With the use of pathway impact analysis, 53 significantly perturbed pathways were identified between the survivor groups. In the second study,[65] the authors further interrogated the hypoxia-inducible factor 1 (HIF-1) signaling pathway that was identified in their previous analysis using both transcriptomic and epigenomic data. Of the 100 genes in the HIF-1 signaling pathway, eight were found to be differentially expressed and methylated between the survivor groups. Next, these eight genes were evaluated in preclinical models of neuropathic pain using publicly available datasets from the National Center for Biotechnology Information Gene Expression Omnibus[66] (ncbi.nlm.nih.gov/geo/). Differential expression and methylation of the mitogen-activated protein kinase I interacting serine/threonine kinase I gene was to be found associated with neuropathic pain in both breast cancer survivors with paclitaxel-induced neuropathy and preclinical models of neuropathic pain. Taken together, these findings highlight the strengths of interdisciplinary collaboration and use of multiple types of data sources (eg, omics, preclinical) and suggest a potential target for intervention.

## Skills Needed to Harness Big Data

Given that big data is increasingly being used to inform clinical practice, it is imperative that nurse scientists have the requisite knowledge and skills to use these data. All nurses should be taught the definition, characteristics, applications, and limitations of big data.[67] Nurse scientists who intend to collect and analyze big data should pursue training in the computational methods described in Papachristou et al's commentary on big data analytics in this Big Data Special Issue. A nonexhaustive list of educational opportunities for nurse scientists who wish to pursue training in the collection or analysis of big data is provided in Table 1. In addition, nurses in all roles should be skilled at interdisciplinary collaboration. Data scientists, bioinformaticians, and computer scientists have the expertise to support nurses to extract, organize, and analyze large datasets. In turn, nurses provide the holistic perspectives required to interpret and act on the results of these analyses to improve the well-being of individuals, families, and communities.[9]

## Ethical Considerations with Big Data

### *Informed Consent*

Informed consent in the context of big data is a subject of significant ethical discourse that is centered on concerns about participant autonomy.[68-70] For example, participants who consent to have their blood collected for a genome-wide association study may not anticipate the discovery of secondary findings related to a pathogenic gene variant, such as for *BRCA1* or *BRCA2*. In addition, they may not be fully prepared to share this information with relatives or future offspring. While a participant may provide a specimen for a candidate gene association study of inflammatory markers, in a case where broad consent is obtained, this specimen may be used for future research (eg, genome-wide study of pathogenic variants). These considerations must be included in the informed consent process to ensure autonomy is upheld. For more detailed information on broad consent in the context of omics research, refer to the excellent review by Williams and Anderson.[71]

In terms of informed consent for studies using social media data, individuals grant specific permissions to social media platforms during registration. However, these permissions are not knowingly extended to recruitment and data collection for research.[72,73] Therefore, researchers need to identify their presence in both public and private social media groups and be transparent in their intentions with potential and recruited participants. In addition, given that assurances of anonymity in social media research cannot be promised, strict procedures to strengthen confidentiality must be made throughout the research process.[73,74]

### *Duty to Report or Intervene*

When accruing, analyzing, or mining big data, procedures must be in place to respond to or intervene on issues of participant safety or to address incidental findings. These considerations are important given that the methods for big data collection and analysis may not facilitate the real-time evaluation of individuals' responses. For example, in clinical trials, the collection of PRO data on emotional distress or pain may identify individuals experiencing severe levels of distress or pain that necessitate a timely response. To identify these patients in real-time, researchers can implement specific PRO thresholds that trigger an alert, identify the individual, and allow researchers or clinicians to intervene in a timely manner.[75] In terms of omics data, secondary findings, such as pathogenic or expected pathogenic variants, may be identified.[76] For example, findings from a study that conducted whole-exome sequencing for 49,960 participants in the UK Biobank reported that 2.7% of participants had a pathogenic or likely pathogenic variant as defined by the American College of Medical Genetics and Genomics Secondary Findings Guidelines.[77] Under the UK Biobank informed consent, these results cannot be shared with participants or their clinicians. In the All of Us Research Program that includes an evaluation of 59 pathogenic or expected pathogenic variants, participants are given the option during the informed consent process to receive this information.[78] In addition, if medically actionable variants are identified, participants will receive genetic counseling.

### *Security and Privacy*

Given the depth and breadth of big data, security of these data is a significant issue that will only magnify as data accrues. For example, data breaches in healthcare systems containing millions of EHRs are not uncommon.[79] Nurse engagement in all steps of the research process is required to ensure that safeguards are in place to protect patient data.

Specific security and privacy concerns apply to data collected from sensors and wearable devices. When third-party technologies are used to collect research data, the amount and type of data that device manufacturers collect from participants are often beyond the investigator's control.[80] Both breaches in data security and increased surveillance have the potential to harm participants by violating their right to privacy. Investigators who collect data using sensors and wearable devices can support participants' right to privacy by including information about how data may be used by third parties in the informed consent document.[80]

Engagement in policy development to ensure patient protections is an important role for oncology nurse scientists who use big data.[81] In addition, nurse clinicians and researchers must have a keen

**TABLE 1**
Educational resources and training opportunities for nurses on use of big data.

| Content Area | Name | Description | Location of Course/Training |
|---|---|---|---|
| Data Science Post-doctoral Fellowship | Big Data Scientist Training Enhancement Program | • Two-year fellowship offered through the National Cancer Institute and Veterans Health Administration of the USA<br>• Goal is to leverage data science to advance cancer research through training, clinical guidance, and use of the VA health data infrastructure | Onsite at one of four VA medical centers<br>Applicants must be a citizen of the USA; have proficiency in at least one programming language, and have experience in bioinformatics, modeling, or management of large datasets |
| • EMR<br>• Statistical Methods | Electronic Medical Records Boot Camp: Biostatistical Methods for Analyzing EMR Data | • Two-day intensive boot camp featuring hands-on training and seminars<br>• Provides an overview of electronic health data opportunities, statistical challenges, and latest techniques | Columbia University SHARP Training Program, New York, NY, USA<br>Offered online |
| • Omics | Big Data Training for Cancer Research | • 10-day intensive workshop geared toward cancer researchers<br>• Workshop covers managing, visualizing, analyzing, and integrating various types of omics data | Purdue University, West Lafayette, IN, USA<br>Offered online or in-person |
| • Genomics | Computational Genomics | • Seven-day intensive course focused on the theory and practice of algorithms in computational biology<br>• Course topics include statistical considerations in the design and analysis of genomic experiments | Cold Springs Harbor Laboratory<br>Offered online |
| • Genomics | Quantitative Genomics Training: Methods and tools for whole-genome and transcriptome analyses | • Two-day intensive boot camp featuring hands-on training and seminars<br>• Provides an overview of concepts, methods, and tools for whole-genome and transcriptome analyses in human health studies | Columbia University SHARP Training Program, New York, NY, USA<br>Offered online |
| • Epigenetics | Epigenetics Boot Camp: Planning and Analyzing DNA Methylation Studies for Human Populations | • Two-day intensive boot camp featuring hands-on training and seminars<br>• Provides overview of concepts, techniques, and data analysis methods utilized in human epigenetic studies with a focus on DNA methylation array | Columbia University SHARP Training Program, New York, NY, USA<br>Offered online |
| • Microbiome | Microbiome Data Analytics Boot Camp: Planning, generating, and analyzing 16s rRNA gene sequencing surveys | • Two-day intensive boot camp featuring hands-on training and seminars<br>• Provides an overview of 16s rRNA gene sequencing surveys including planning, generating, and analyzing sequencing datasets | Columbia University SHARP Training Program, New York, NY, USA<br>Offered online |
| • Microbiome | Strategies and Techniques for Analyzing Microbial Population Structures (STAMPS) | • 10-day intensive course providing interdisciplinary training in bioinformatics and statistics<br>• Topics include experimental design, assembly and annotation of shotgun metagenomic data, and statistical methods | Marine Biological Laboratory, Woods Hole, MA, USA<br>Offered in-person |
| • Statistical methods<br>• R programming<br>• Omics | University of Washington Biostatistics Summer Institute in Statistical Genomics (SISG) | • SISG is divided into several modules scheduled throughout summer<br>• Provides instruction on the modern methods of statistical analysis | University of Washington, Seattle, WA, USA<br>Offered in-person |
| • Statistical methods | University of Washington Summer Institute in Statistics for Big Data | • Four modules provided over 2 to 3 days<br>• Introduces statistical techniques for the analysis of biological big data | University of Washington, Seattle, WA, USA<br>Offered online |
| • Machine learning | Machine Learning Boot Camp: Analyzing Biomedical and Health Data | • Two-day intensive boot camp featuring hands-on training and seminars<br>• Provides overview of statistical concepts, techniques, and data analysis methods for biomedical research | Columbia University SHARP Training Program, New York, NY, USA<br>Offered in-person or online |
| • Machine learning | Introduction to Machine Learning for Epidemiologists | • Course is a month-long, 30-hour, self-paced digital course<br>• Provides a broad exposure to machine learning and its practical applications within epidemiology | episummer, Columbia University, New York, NY, USA<br>Offered online |
| • R programming<br>• Python<br>• Machine Learning<br>• Text Mining | Bioinformatics Course Series and Workshops offered through FAES Academic Programs at the National Institutes of Health | • Courses are offered online in an asynchronous format<br>• Workshops are offered online in a synchronous format | FAES Academic Programs, Bethesda, MD, USA<br>Offered online |
| • R programming<br>• Bioinformatics<br>• Machine learning | Coursera (coursera.org) | • Self-paced online courses covering a variety of topics and levels of proficiency | Offered online |
| • R programming<br>• Python<br>• Machine learning | DataCamp (datacamp.com) | • Self-paced online courses covering a variety of topics and levels of proficiency | Offered online |

Abbreviations: EMR = electronic medical record, FAES = Foundation for Advanced Education in the Sciences, SHARP = Skills for Health and Research Professionals, SISG = Summer Institute in Statistical Genomics, USA = United States of America, VA = Veterans Affairs

knowledge of the policies that regulate big data and the limitations of these policies to ensure that all facets of the polices are adhered to and to serve as a resource to patients. One example of policy that seeks to regulate big data is the General Data Protection Regulation of the European Union. Effective since 2018, this law restricts how any entity, within or outside of the European Union, may handle or process personal data of citizens or residents of the European Union.[82] Reinforced with steep fines, this law outlines the rights of the data subject (eg, right to restriction of processing), rules of consent, conditions when personal data may be processed, responsibilities of data controllers and processors, and expectations for data protection.

In terms of genetic data, the Genetic Information and Nondiscrimination Act (GINA) was passed in the United States to protect individuals who provide their genetic information for research studies from the potential for genetic discrimination in terms of employment and health insurance.[83] Specifically, employers cannot discriminate in terms of hiring or firing an individual based on their genetic information and cannot request this information from employees. In addition, health insurers cannot deny coverage or change insurance rates based on an individual's genetic information. Genetic information in these instances extend beyond the individual and include family members. However, GINA does not protect individuals from genetic discrimination in terms of life insurance, disability insurance, long-term care insurance, or other uses of genetic information.[84] Furthermore, GINA only applies to individuals who have not been diagnosed with a medical condition associated with their genetic makeup. Therefore, this law does not apply to cancer survivors. Similar laws were implemented in Canada (ie, the Genetic Non-Discrimination Act)[85] and Germany (ie, German Human Genetic Examination Act).[86] For ongoing discussion on the ethical, legal, and social implications of genomics research, refer to the review by Hammer.[87]

### *Underrepresentation in Big Data*

As with other types of research, the underrepresentation of individuals from minoritized racial, ethnic, sexual, and gender groups in big data delays progress toward precision health[74] and can lead to harmful study findings and/or interpretation. For example, in a study that examined the ancestral population diversity in two public data sources from the NIH (ie, Genome-Wide Association Study Catalog, dbGaP), African, Latin American, and Asian ancestral populations were significantly underrepresented.[88] In genomic research, underrepresentation of these ancestral populations in diverse datasets may hinder the identification of gene–disease associations that are uncommon in European ancestral populations, lead to the identification of incorrect associations, and limit the generalizability of findings in the clinical setting.

Underrepresentation in big data is particularly problematic when these data are used to train machine learning models. For example, lack of racial and ethnic diversity in publicly available radiology datasets has limited the ability of artificial intelligence programs to correctly identify breast lesions in patients of color.[89,90] To address this issue, a team of researchers from Emory University in the United States developed the EMory BrEast imaging Dataset (EMBED), which includes detailed demographic, lesion, and pathological data on a diverse sample of nearly 116,000 patients.[90] The researchers hypothesize that this diverse dataset will allow for the "development and validation of deep learning models for breast cancer screening that perform equally across patient demographic characteristics and reduce disparities in health care" (p. 7).[90] Importantly, underrepresentation is not the only source of potential bias in research that uses big data.[91,92] Investigators have a responsibility to familiarize themselves with the principles of algorithmic fairness and the potential for latent biases to influence the results of big data studies.

## Future Directions and Conclusion

The authors summarized the state of the science of big data in oncology nursing research by describing common sources of big data, reviewing access considerations to these data sources, and providing exemplars on how these sources can be used to examine research questions relevant to oncology nursing research. While the emergence of big data and advances in analytic approaches provide new and exciting opportunities to advance oncology nursing science, they pose several challenges for nurse clinicians and researchers. For nurse clinicians, these challenges may include the facilitation of data collection from remote devices, staying current of rapidly evolving genomic tests to provide patient education and support,[87] and translating findings from big data analyses into practice. Nurse researchers require education and training to develop research questions using and surmount challenges associated with rapidly evolving data analytic methods. For both nurse clinicians and researchers, ethical challenges associated with big data are ongoing and are likely to become more prominent with the increasingly ubiquitous nature of passive data collection. Common to each of these challenges is the need for education. As stated previously, all nurses need to understand big data, both its applications and limitations. Nursing programs need to provide courses on big data at all levels that include discussions of ethics and statistical methods. Nurses who are trained in big data and advanced computational methods will be poised to contribute to guidelines and policies that preserve the rights of human research participants. Big data has the potential to provide a current, comprehensive, and holistic representation of the patient's experience. With their unique focus on the interplay between the whole person, the environment, and health, nurses bring an indispensable perspective to the interpretation and application of big data research findings. Using these approaches, oncology nurses will stay on the forefront of advancements in big data approaches and harness big data to improve the outcomes of patients with cancer.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this report.

## Funding

Dr. Miaskowski is an American Cancer Society Clinical Research Professor. Dr. Harris is supported by a grant from the National Institute of Nursing Research of the National Institutes of Health (NR009759). Dr. Kober is partially supported by a grant from the National Cancer Institute of the National Institutes of Health (CA233774). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Carolyn S. Harris reports financial support was provided by National Institute of Nursing Research.

## References

1. Definition of Big Data - Gartner Information Technology Glossary. Gartner. Accessed February 15, 2023. https://www.gartner.com/en/information-technology/glossary/big-data.
2. Curry E. The big data value chain: definitions, concepts, and theoretical approaches. In: Cavanillas JM, Curry E, Wahlster W, eds. *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe*. Springer International Publishing; 2016:29–37.
3. Pablo R-GJ, Roberto D-P, Victor S-U, Isabel G-R, Paul C, Elizabeth O-R. Big data in the healthcare system: a synergy with artificial intelligence and blockchain technology. *J Integr Bioinform*. 2021;19(1). https://doi.org/10.1515/jib-2020-0035.
4. Ristevski B, Chen M. Big data analytics in medicine and healthcare. *J Integr Bioinform*. 2018;15(3). https://doi.org/10.1515/jib-2017-0030.

5. Viceconti M, Hunter P, Hose R. Big data, big knowledge: big data for personalized healthcare. *IEEE J Biomed Health Inform*. 2015;19(4):1209–1215. https://doi.org/10.1109/JBHI.2015.2406883.
6. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. *J Big Data*. 2019;6(1):1–25. https://doi.org/10.1186/s40537-019-0217-0.
7. Office of the National Coordinator for Health Information Technology. What are patient-generated health data? Accessed February 15, 2023. https://www.healthit.gov/topic/otherhot-topics/what-are-patient-generated-health-data.
8. Broniatowski DA, Dredze M, Paul MJ, Dugas A. Using social media to perform local influenza surveillance in an inner-city hospital: a retrospective observational study. *JMIR Public Health Surveill*. 2015;1(1):e5. https://doi.org/10.2196/publichealth.4472.
9. Brennan PF, Bakken S. Nursing needs big data and big data needs nursing. *J Nurs Scholarsh*. 2015;47(5):477–484. https://doi.org/10.1111/jnu.12159. 2015/9.
10. Aiken LH, Clarke SP, Sloane DM, Sochalski J, Silber JH. Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *JAMA*. 2002;288(16):1987–1993. https://doi.org/10.1001/jama.288.16.1987.
11. Bergstrom N, Braden B, Kemp M, Champagne M, Ruby E. Multi-site study of incidence of pressure ulcers and the relationship between risk level, demographic characteristics, diagnoses, and prescription of preventive interventions. *J Am Geriatr Soc*. 1996;44(1):22–30. https://doi.org/10.1111/j.1532-5415.1996.tb05633.x. 1996/1.
12. Wu MW, Lee TT, Lai SM, Huang CY, Chang TH. Evaluation of electronic health records on the nursing process and patient outcomes regarding fall and pressure injuries. *Comput Inform Nurs*. 2019;37(11):573–582. https://doi.org/10.1097/cin.0000000000000548.
13. Durieux BN, Tarbi EC, Lindvall C. Opportunities for computational tools in palliative care: supporting patient needs and lowering burden. *Palliat Med*. 2022;36(8):1168–1170. https://doi.org/10.1177/02692163221122261.
14. Westra BL, Clancy TR, Sensmeier J, Warren JJ, Weaver C, Delaney CW. Nursing knowledge: big data science-implications for nurse leaders. *Nurs Adm Q*. 2015;39(4):304–310. https://doi.org/10.1097/NAQ.0000000000000130.
15. O'Brien RL, O'Brien MW. CE: nursing orientation to data science and machine learning. *Am J Nurs*. 2021;121(4):32–39. https://doi.org/10.1097/01.NAJ.0000742064.59610.28.
16. Fu MR, Kurnat-Thoma E, Starkweather A, et al. Precision health: a nursing perspective. *Int J Nurs Sci*. 2020;7(1):5–12. https://doi.org/10.1016/j.ijnss.2019.12.008.
17. Masukawa K, Aoyama M, Yokota S, et al. Machine learning models to detect social distress, spiritual pain, and severe physical psychological symptoms in terminally ill patients with cancer from unstructured text data in electronic medical records. *Palliat Med*. 2022;36(8):1207–1216. https://doi.org/10.1177/02692163221105595.
18. Kreimeyer K, Foster M, Pandey A, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *J Biomed Inform*. 2017;73:14–29. https://doi.org/10.1016/j.jbi.2017.07.012.
19. Koleck TA, Tatonetti NP, Bakken S, et al. Identifying symptom information in clinical notes using natural language processing. *Nurs Res*. 2021;70(3):173–183. https://doi.org/10.1097/nnr.0000000000000488.
20. Koleck TA, Topaz M, Tatonetti NP, et al. Characterizing shared and distinct symptom clusters in common chronic conditions through natural language processing of nursing notes. *Res Nurs Health*. 2021;44(6):906–919. https://doi.org/10.1002/nur.22190.
21. Lindvall C, Deng C-Y, Moseley E, et al. Natural language processing to identify advance care planning documentation in a multisite pragmatic clinical trial. *J Pain Symptom Manage*. 2022;63(1):e29–e36. https://doi.org/10.1016/j.jpainsymman.2021.06.025. PMID - 34271146.
22. Deering MJ. *Issue Brief: Patient-Generated Health Data and Health IT*. The Office of the National Coordinator for Health Information Technology; December 20, 2013. Published; https://www.healthit.gov/sites/default/files/pghd_brief_final122013.pdf.
23. Pratap A, Renn BN, Volponi J, et al. Using mobile apps to assess and treat depression in Hispanic and Latino populations: fully remote randomized clinical trial. *J Med Internet Res*. 2018;20(8):e10130. https://doi.org/10.2196/10130.
24. Zhang Y, Pratap A, Folarin AA, et al. Long-term participant retention and engagement patterns in an app and wearable-based multinational remote digital depression study. *NPJ Digit Med*. 2023;6(1):25. https://doi.org/10.1038/s41746-023-00749-3.
25. Kofoed S, Breen S, Gough K, Aranda S. Benefits of remote real-time side-effect monitoring systems for patients receiving cancer treatment. *Oncol Rev*. 2012;6(1):e7. https://doi.org/10.4081/oncol.2012.e7.
26. Cox SM, Lane A, Volchenboum SL. Use of wearable, mobile, and sensor technology in cancer clinical trials. *JCO Clin Cancer Inform*. 2018;2:1–11. https://doi.org/10.1200/CCI.17.00147.
27. Wood WA, Bennett AV, Basch E. Emerging uses of patient generated health data in clinical research. *Mol Oncol*. 2015;9(5):1018–1024. https://doi.org/10.1016/j.molonc.2014.08.006.
28. Kvedar J, Coye MJ, Everett W. Connected health: a review of technologies and strategies to improve patient care with telemedicine and telehealth. *Health Aff*. 2014;33(2):194–199. https://doi.org/10.1377/hlthaff.2013.0992.
29. Dias D, Paulo Silva Cunha J. Wearable health devices:vital sign monitoring, systems and technologies. *Sensors*. 2018;18(8). https://doi.org/10.3390/s18082414.
30. Innominato PF, Komarzynski S, Mohammad-Djafari A, et al. Clinical relevance of the first domomedicine platform securing multidrug chronotherapy delivery in metastatic cancer patients at home: the inCASA European Project. *J Med Internet Res*. 2016;18(11):e305. https://doi.org/10.2196/jmir.6303.
31. Innominato P, Komarzynski S, Karaboué A, et al. Home-based e-health platform for multidimensional telemonitoring of symptoms, body weight, sleep, and circadian activity: relevance for chronomodulated administration of irinotecan, fluorouracil-leucovorin, and oxaliplatin at home-results from a pilot study. *JCO Clin Cancer Inform*. 2018;2:1–15. https://doi.org/10.1200/CCI.17.00125.
32. Holländer-Mieritz C, Vogelius IR, Kristensen CA, Green A, Rindum JL, Pappot H. Using biometric sensor data to monitor cancer patients during radiotherapy: protocol for the OncoWatch Feasibility Study. *JMIR Res Protoc*. 2021;10(5):e26096. https://doi.org/10.2196/26096.
33. LeBaron V, Alam R, Bennett R, et al. Deploying the behavioral and environmental sensing and intervention for cancer smart health system to support patients and family caregivers in managing pain: feasibility and acceptability study. *JMIR Cancer*. 2022;8(3):e36879. https://doi.org/10.2196/36879.
34. Kotronoulas G, Kearney N, Maguire R, et al. What is the value of the routine use of patient-reported outcome measures toward improvement of patient outcomes, processes of care, and health service outcomes in cancer care? A systematic review of controlled trials. *J Clin Oncol*. 2014;32(14):1480–1501. https://doi.org/10.1200/jco.2013.53.5948.
35. Calvert M, Thwaites R, Kyte D, Devlin N. Putting patient-reported outcomes on the 'Big Data Road Map. *J R Soc Med*. 2015;108(8):299–303. https://doi.org/10.1177/0141076815579896.
36. Austin E, LeRouge C, Hartzler AL, Segal C, Lavallee DC. Capturing the patient voice: implementing patient-reported outcomes across the health system. *Qual Life Res*. 2020;29(2):347–355. https://doi.org/10.1007/s11136-019-02320-8.
37. Hsiao C-J, Dymek C, Kim B, Russell B. Advancing the use of patient-reported outcomes in practice: understanding challenges, opportunities, and the potential of health information technology. *Qual Life Res*. 2019;28(6):1575–1583. https://doi.org/10.1007/s11136-019-02112-0.
38. Crown W, Clancy TR. Working in the New Big Data World: Academic/Corporate Partnership Model. In Delaney CW, Weaver CA, Warren JJ, Clancy TR, Simpson RL (Eds.), Big data-enabled nursing: Education, research and practice. (pp. 157-171). Cham, Switzerland: Springer International.
39. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc*. 2014;21(4):578–582. https://doi.org/10.1136/amiajnl-2014-002747.
40. van de Poll-Franse LV, Horevoorts N, Schoormans D, et al. Measuring clinical, biological, and behavioral variables to elucidate trajectories of patient-reported outcomes: the PROFILES Registry. *J Natl Cancer Inst*. 2022;114(6):800–807. https://doi.org/10.1093/jnci/djac047.
41. Vaz-Luis I, Cottu P, Mesleard C, et al. UNICANCER: French prospective cohort study of treatment-related chronic toxicity in women with localised breast cancer (CANTO). *ESMO Open*. 2019;4(5): e000562. https://doi.org/10.1136/esmoopen-2019-000562.
42. Organization for Economic Cooperation and Development. Patient-Reported Indicator Surveys (PaRIS). Accessed March 7, 2023. https://www.oecd.org/health/paris/.
43. Patient Centered Outcomes For Health Measures. ICHOM. Published July 27, 2022. Accessed Mar 8, 2023. https://www.ichom.org/patient-centered-outcome-measures/.
44. National Institutes of Health. Final NIH Policy for Data Management and Sharing. Accessed Mar 6, 2023. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html.
45. UK Biobank Coordinating Center. About us. Accessed March 15, 2023. https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us.
46. UK Biobank Coordinating Center. UK Biobank: Protocol for a large-scale prospective epidemiological resource. Published 2007. Accessed February 15, 2023. https://www.ukbiobank.ac.uk/media/gnkeyh2q/study-rationale.pdf.
47. National Institutes of Health. All of Us Research Program Overview. Published June 22, 2020. Accessed March 15, 2023. https://allofus.nih.gov/about/program-overview.
48. Precision Medicine Initiative Working Group. The Precision Medicine Initiative Cohort Program - Building a Research Foundation for the 21st Century. NIH All of Us Research Program; 2015. Accessed February 2, 2023. www.nih.gov/sites/default/files/research-training/initiatives/pmi/pmi-working-group-report-20150917-2.pdf.
49. Dreisbach C, Grayson S, Leggio K, Conway A, Koleck T. Predictors of unrelieved symptoms in All of Us Research Program participants with chronic conditions. *J Pain Symptom Manage*. 2022;64(6):555–566. https://doi.org/10.1016/j.jpainsymman.2022.08.018.
50. Agency for Healthcare Research and Quality. What Is Patient Experience? Accessed March 15, 2023. https://www.ahrq.gov/cahps/about-cahps/patient-experience/index.html.
51. Agency for Healthcare Research and Quality. CAHPS Cancer Care Survey. Accessed March 15, 2023. https://www.ahrq.gov/cahps/surveys-guidance/cancer/index.html.
52. National Health Service England. National Cancer Patient Experience Survey. Published May 22, 2020. Accessed March 15, 2023. https://www.ncpes.co.uk/.
53. Fauer A, Choi SW, Wallner LP, Davis MA, Friese CR. Understanding quality and equity: patient experiences with care in older adults diagnosed with hematologic malignancies. *Cancer Causes Control*. 2021;32(4):379–389. https://doi.org/10.1007/s10552-021-01395-4.
54. Alessy SA, Davies E, Rawlinson J, Baker M, Lüchtenborg M. Clinical nurse specialists and survival in patients with cancer: the UK National Cancer Experience Survey. *BMJ Support Palliat Care*. 2022. https://doi.org/10.1136/bmjspcare-2021-003445.
55. Arditi C, Eicher M, Colomer-Lahiguera S, et al. Patients' experiences with cancer care in Switzerland: results of a multicentre cross-sectional survey. *Eur J Cancer Care*. 2022;31(6):e13705. https://doi.org/10.1111/ecc.13705.

56. Kemp S. Digital 2023: Global Overview Report. DataReportal – Global Digital Insights. Published January 26, 2023. Accessed March 15, 2023. https://datareportal.com/reports/digital-2023-global-overview-report.
57. Allen KA, Jimerson SR, Quintana DS, McKinley L. *An Academic's Guide to Social Media: Learn, Engage, and Belong*. Routledge; 2022.
58. Pew Research Center: Internet, Science & Tech. Social Media Fact Sheet. Published April 7, 2021. Accessed February 10, 2023. https://www.pewresearch.org/internet/fact-sheet/social-media/.
59. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. *J Med Internet Res*. 2013;15(4):e85. https://doi.org/10.2196/jmir.1933.
60. Pozzar R, Hammer MJ, Underhill-Blazey M, et al. Threats of bots and other bad actors to data quality following research participant recruitment through social media: cross-sectional questionnaire. *J Med Internet Res*. 2020;22(10):e23021. https://doi.org/10.2196/23021.
61. Lee YJ, Jang H, Campbell G, Carenini G, Thomas T, Donovan H. Identifying language features associated with needs of ovarian cancer patients and caregivers using social media. *Cancer Nurs*. 2022;45(3):E639–e645. https://doi.org/10.1097/ncc.0000000000000928.
62. Harris CS, Miaskowski CA, Dhruva AA, Cataldo J, Kober KM. Multi-staged data-integrated multi-omics analysis for symptom science research. *Biol Res Nurs*. 2021;23(4):596–607. https://doi.org/10.1177/10998004211003980.
63. Ray M, Wallace MK, Grayson SC, et al. Epigenomic links between social determinants of health and symptoms: a scoping review. *Biol Res Nurs*. 2022. https://doi.org/10.1177/10998004221147300.
64. Kober KM, Olshen A, Conley YP, et al. Expression of mitochondrial dysfunction-related genes and pathways in paclitaxel-induced peripheral neuropathy in breast cancer survivors. *Molecular Pain*. 2018;14:1–16. https://doi.org/10.1177/1744806918816462.
65. Kober KM, Lee MC, Olshen A, et al. Differential methylation and expression of genes in the hypoxia-inducible factor 1 signaling pathway are associated with paclitaxel-induced peripheral neuropathy in breast cancer survivors and with preclinical models of chemotherapy-induced neuropathic pain. *Molecular Pain*. 2020;16:1–15. https://doi.org/10.1177/1744806920936502.
66. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res*. 2012;41(D1):D991–D995. https://doi.org/10.1093/nar/gks1193.
67. Topaz M, Pruinelli L. Big data and nursing: implications for the future. *Stud Health Technol Inform*. 2017;232:165–171.
68. Fiore RN, Goodman KW. Precision medicine ethics: selected issues and developments in next-generation sequencing, clinical oncology, and ethics. *Curr Opin Oncol*. 2016;28(1):83–87. https://doi.org/10.1097/CCO.0000000000000247.
69. Howe EGI, Elenberg F. Ethical Challenges Posed by Big Data. *Innov Clin Neurosci*. 2020;17(10-12):24–30.
70. Smilan LE. Broad Consent-Are We Asking Enough? *Ethics Hum Res*. 2022;44(5):22–31. https://doi.org/10.1002/eahr.500140.
71. Williams JK, Anderson CM. Omics research ethics considerations. *Nurs Outlook*. 2018;66(4):386–393. https://doi.org/10.1016/j.outlook.2018.05.003.
72. Hunter RF, Gough A, O'Kane N, et al. Ethical issues in social media research for public health. *Am J Public Health*. 2018;108(3):343–348. https://doi.org/10.2105/AJPH.2017.304249.
73. Kamp K, Herbell K, Magginis WH, Berry D, Given B. Facebook recruitment and the protection of human subjects. *West J Nurs Res*. 2019;41(9):1270–1281. https://doi.org/10.1177/0193945919828108.
74. Hammer MJ, Pozzar R. And (in)justice for all: disparities in oncology research. 2020;11(1):95-107. https://doi.org/10.1615/EthicsBiologyEngMed.2021038584.
75. Kyte D, Ives J, Draper H, Keeley T, Calvert M. Inconsistencies in quality of life data collection in clinical trials: a potential source of bias? Interviews with research nurses and trialists. *PLoS One*. 2013;8(10):e76625. https://doi.org/10.1371/journal.pone.0076625.
76. Miller DT, Lee K, Gordon AS, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2021 update: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet Med*. 2021;23(8):1391–1398. https://doi.org/10.1038/s41436-021-01171-4.
77. Van Hout CV, Tachmazidou I, Backman JD, et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature*. 2020;586(7831):749–756. https://doi.org/10.1038/s41586-020-2853-0.
78. All of Us Research Program. *NIH's All of Us Research Program returns genetic health-related results to participants*. 2022. https://allofus.nih.gov/news-events/announcements/nihs-all-us-research-program-returns-genetic-health-related-results-participants.
79. Healthcare Data Breach Statistics. HIPAA Journal. Published February 21, 2023. Accessed March 15, 2023. https://www.hipaajournal.com/healthcare-data-breach-statistics/.
80. Sui A, Sui W, Liu S, Rhodes R. Ethical considerations for the use of consumer wearables in health research. *Digit Health*. 2023;9:20552076231153740. https://doi.org/10.1177/20552076231153740.
81. Hickey KT, Bakken S, Byrne MW, et al. Precision health: advancing symptom and self-management science. *Nurs Outlook*. 2019;67(4):462–475. https://doi.org/10.1016/j.outlook.2019.01.003.
82. L119 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Official Journal of the European Union) 2016:1-88.
83. Rothstein MA. Putting the Genetic Information Nondiscrimination Act in context. *Genet Med*. 2008;10(9):655–656. https://doi.org/10.1097/gim.0b013e31818337bd.
84. Joly Y, Dupras C, Pinkesz M, Tovino SA, Rothstein MA. Looking Beyond GINA: Policy Approaches to Address Genetic Discrimination. *Annu Rev Genomics Hum Genet*. 2020;21:491–507. https://doi.org/10.1146/annurev-genom-111119-011436.
85. Government of Canada. Genetic Non-Discrimination Act. Published May 4, 2017. Accessed March 15, 2023. https://laws-lois.justice.gc.ca/eng/acts/G-2.5/page-1.html.
86. German Federal Parliament. Human Genetic Examination Act. European Society for Human Genetics. Published 2009. Accessed February 16, 2023. https://www.eshg.org/fileadmin/www.eshg.org/documents/Europe/LegalWS/Germany_GenDG_Law_German_English.pdf.
87. Hammer MJ. Beyond the helix: Ethical, legal, and social implications in genomics. *Semin Oncol Nurs*. 2019;35(1):93–106. https://doi.org/10.1016/j.soncn.2018.12.007.
88. Landry LG, Ali N, Williams DR, Rehm HL, Bonham VL. Lack of diversity in genomic databases is a barrier to translating precision medicine research into practice. *Health Aff*. 2018;37(5):780–785. https://doi.org/10.1377/hlthaff.2017.1595.
89. Thrall JH, Li X, Li Q, et al. Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. *J Am Coll Radiol*. 2018;15(3):504–508. https://doi.org/10.1016/j.jacr.2017.12.026. Pt B.
90. Jeong JJ, Vey BL, Bhimireddy A, et al. The EMory BrEast imaging Dataset (EMBED): a racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. *Radiol Artif Intell*. 2023;5(1): e220047. https://doi.org/10.1148/ryai.220047.
91. DeCamp M, Lindvall C. Latent bias and the implementation of artificial intelligence in medicine. *J Am Med Inform Assoc*. 2020;27(12):2020–2023. https://doi.org/10.1093/jamia/ocaa094.
92. Al-Zaiti SS, Alghwiri AA, Hu X, et al. A clinician's guide to understanding and critically appraising machine learning studies: a checklist for Ruling Out Bias Using Standard Tools in Machine Learning (ROBUST-ML). *Eur Heart J Digit Health*. 2022;3(2):125–140. https://doi.org/10.1093/ehjdh/ztac016.