special article

Susan G. Komen Big Data for Breast Cancer Initiative: How Patient Advocacy Organizations Can Facilitate Using Big Data to Improve Patient Outcomes

Jerome Jourquin, PhD, MS[1]; Stephanie Birkey Reffey, PhD[1]; Cheryl Jernigan[2]; Mia Levy, MD, PhD[3]; Glendon Zinser, PhD[1]; Kimberly Sabelko, PhD[1]; Jennifer Pietenpol, PhD[4]; and George Sledge Jr, MD[5]

abstract

Integrating different types of data, including electronic health records, imaging data, administrative and claims databases, large data repositories, the Internet of Things, genomics, and other omics data, is both a challenge and an opportunity that must be tackled head on. We explore some of the challenges and opportunities in optimizing data integration to accelerate breast cancer discovery and improve patient outcomes. Susan G. Komen convened three meetings (2015, 2017, and 2018) with various stakeholders to discuss challenges, opportunities, and next steps to enhance the use of big data in the field of breast cancer. Meeting participants agreed that big data approaches can enhance the identification of better therapies, improve outcomes, reduce disparities, and optimize precision medicine. One challenge is that databases must be shared, linked with each other, standardized, and interoperable. Patients want to be active participants in research and their own care, and to control how their data are used. Many patients have privacy concerns and do not understand how sharing their data can help to effectively drive discovery. Public education is essential, and breast cancer researchers who are skilled in using and analyzing big data are needed. Patient advocacy groups can play multiple roles to help maximize and leverage big data to better serve patients. Komen is committed to educating patients on big data issues, encouraging data sharing by all stakeholders, assisting in training the next generation of data science breast cancer researchers, and funding research projects that will use real-life data in real time to revolutionize the way breast cancer is understood and treated.

INTRODUCTION

Susan G. Komen envisions a world with a seamless web of health care information, where patients are informed and empowered to use their data to improve their health care; electronic health records (EHRs) are connected to other data sources to provide evidence-based support for clinical decision making; many, if not all, patients participate in clinical research; data systems are linked, secure, and easily accessible; genomics and other omics are universally available and user friendly; researchers can mine enhanced data sets to address questions; and most importantly, fewer people die of breast cancer and quality of life improves for those living with the disease. Unfortunately, this world does not yet exist. The health care community currently faces many challenges and opportunities to efficiently use data to more effectively treat patients. Komen convened three Big Data for Breast Cancer (BD4BC) meetings to foster open dialog among experts and strategically invited a wide array of participants (patient advocates, oncologists, bioethicists, laboratory researchers, genomic- and proteomics-based companies, big data-focused pharmaceutical companies, and data software companies) to discuss the status of, challenges in, and barriers to optimizing big data and the opportunities big data can provide to advance health care (Fig 1).

The first meeting (New York, NY, October 2015) focused on the barriers, opportunities, needs, priorities, and solutions concerning the challenge of improving breast cancer research and care through big data. The follow-up meetings (Menlo Park, CA, February 2017 and 2018) focused on data infrastructure; research in and with big data; clinical applications for big data; and how to leverage data to attain health equity and improve methods for aggregating and analyzing clinical, genomic, and other sources of data for patients with metastatic breast cancer.

Meeting participants (Table 1) defined big data as the integration of large amounts of different types of data, including EHRs, administrative and health insurance

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on July 19, 2019 and published at ascopubs.org/journal/po on September 12, 2019: DOI https://doi.org/10.1200/PO.19.00184

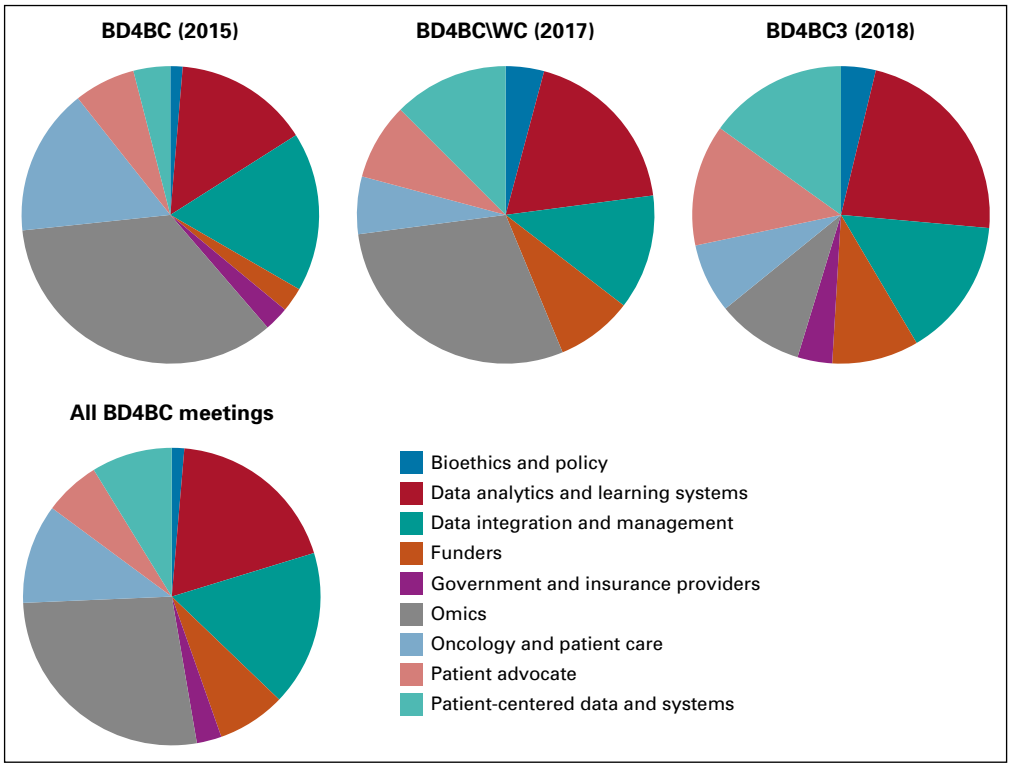

FIG 1. Distribution of Komen's Big Data for Breast Cancer (BD4BC) meeting participants on the basis of their primary expertise. Omics experts were the most represented in the first two meetings (35% and 29%, respectively). A more equal distribution of expertise was reached at the 2018 meeting among patient-centered data and systems, data analytics and learning systems, and data integration and management (15% to 23% each), highlighting the intent to work on solutions to big data challenges in breast cancer and implementation. Looking at all 148 unique participants who attended Komen's BD4BC meetings, the most represented areas of expertise were omics (27%), data analytics and learning systems (19%), data integration and management (17%), and oncology and patient care (11%). BD4BC\WC, West Coast (second BD4BC meeting); BD4BC3, third BD4BC meeting.

claims databases, large data repositories (registries and cohorts), and genomics and other omics data. Participants identified important issues in effectively using big data to improve breast cancer research and clinical care, and they outlined an action plan to focus Komen's efforts. We share the insights gained from these meetings and Komen's planned actions to leverage big data to reduce the number of breast cancer deaths.

As recently as 15 years ago, individual patient data were only available in paper charts located at a single institution and inaccessible to others outside that institution. Data from epidemiologic studies, cooperative group trials, and individual laboratories were often quantitatively modest, dispersed, and not open to sharing except through manuscripts and public presentations.[1] The world of data is rapidly changing, and the amount of data related to breast cancer has exploded. Within the past decade, various types of data are now routinely collected: personal and family history; breast density; patient-reported outcomes; imaging data; clinical trial data; genomics and other omics; annotated mutations; biospecimens; and social, environmental, and behavioral data. Big data can potentially alter how breast cancer is perceived, studied, and treated (Fig 2).

TECHNOLOGY INNOVATION

Technologic innovation has drastically changed the landscape of almost all industries in recent history, including health care, with new technologic advancements made every year. Adoption of EHRs in the United States has profoundly changed how patient data are collected, stored, and used.[2] The scale of data collection through EHRs is stunning. One of us (M.L.) reported at the third BD4BC meeting that in 2015, one health care system's EHR was accessed by 11,000 people per day, creating 6.8 million clinic notes that year, and was associated with 1.7 million outpatient electronic prescriptions and 15 million clinical communications that year.

Furthermore, genomics, proteomics, metabolomics, radiomics, and other emerging omics platforms are expanding the data generated in health care.[3] Decreasing costs and increasing availability of such tests allow clinicians to evaluate tumors via whole exome, genome, or deep sequencing. These are becoming part of the standard of care. Another

TABLE 1. List of Planning Committee Members and Invited Speakers to Each of Komen's BD4BC Meetings

BD4BC New York, NY October 8-9, 2015	BD4BC\WC Menlo Park, CA February 23-24, 2017	BD4BC3 Menlo Park, CA February 1-2, 2018
Sir John Bell, GBE, FRS, FMedSci, FREng (Oxford University)	**Amy Abernethy, MD, PhD (Flatiron Health)**	**Amy Abernethy, MD, PhD (Flatiron Health)**
Nancy Brinker (Susan G. Komen)	**Cheryl Jernigan, CPA, FACHE (Susan G. Komen [patient advocate])**	Regina Barzilay, PhD (Massachusetts Institute of Technology)
Sir Rory Collins, FMedSci FRS (University of Oxford)	**Mia Levy, MD, PhD (Vanderbilt-Ingram Cancer Center)**	Christopher Boone, PhD, FACHE (Pfizer)
Robert Cook-Deegan, MD (Duke Global Health Institute)	Gaurav Singal, MD (Foundation Medicine)	Aradhana Ghosh, MD (Syapse)
Henry Friedman, MD (Duke University)	**George Sledge Jr, MD (Stanford University)**	**Cheryl Jernigan, CPA, FACHE (Susan G. Komen [patient advocate])**
Elad Gil, PhD (Color Genomics)	Gary Thompson, JD, MBA (CLOUD)	Gaurav Kaushik, PhD (Foundation Medicine)
Todd Golub, MD (Dana-Farber Cancer Institute)	**Crystal Valentine, PhD (MapR Technologies)**	**Mia Levy, MD, PhD (Vanderbilt-Ingram Cancer Center)**
Cheryl Jernigan, CPA, FACHE (Susan G. Komen [patient advocate])	**Nikhil Wagle, MD (The Broad Institute and Dana-Farber Cancer Institute)**	**Joshua Mann (SHARE for Cures and Inspirata)**
Mia Levy, MD, PhD (Vanderbilt-Ingram Cancer Center)		John Mattison, MD (Kaiser Permanente)
Jane Perlmutter, PhD (patient advocate)		Joan Neuner, MD, MPH (Medical College of Wisconsin)
Judy Salerno, MD, MS (Susan G. Komen)		Lily Peng, MD, PhD (Google Research)
Charles Sawyers, MD (Howard Hughes Medical Institute, Memorial Sloan Kettering Cancer Center)		Jennifer Pietenpol, PhD (Vanderbilt-Ingram Cancer Center)
George Sledge Jr, MD (Stanford University)		Katherine Reeder-Hayes, MD, MBA, MS (University of North Carolina, Chapel Hill)
Marc Tessier-Levigne, FRS, FRSC, FMedSci (Rockefeller University)		**George Sledge Jr, MD (Stanford University)**
Eric Winer, MD (Dana-Farber Cancer Institute)		Shyrea Thompson (Susan G. Komen)

NOTE. Planning Committee members are in bold type. Affiliations are at the time of each meeting.

Abbreviations: BD4BC\WC, West Coast (second BD4BC meeting); BD4BC3, third BD4BC meeting; CLOUD, Consortium for Local Ownership and Use of Data.

area with a rapidly growing platform generating large amounts of individual-specific data is the Internet of Things (IoT), which includes all devices attached to the Internet that generate data.[4] Wearable devices (eg, fitness trackers) generate continuous, multiparameter, individual-specific data. Garments and ingestible devices are emerging technologies.[4] IoT data can be downloaded, summarized across populations, and attached to other data sets for analysis. By 2020, an estimated 200 billion devices will be connected to the Internet and generating continuous data, with approximately 30% of these data predicted to have medical applications.[4]

With rapid technologic advances and vast amounts of data come challenges that must be overcome to make efficient use of the technology and data being collected. Big data technologies, such as EHRs, omics, and IoT, provide great potential, but these technologies remain siloed and not interoperable.[5,6] Data are often limited geographically and by age group and type of care. Worse, data are often subject to misinterpretation regarding the intent when initially collected,[7] partly because of lack of annotation and documentation. The use of these technologies is also limited by the inability of different platforms to communicate with each other, hospital firewalls that create barriers to sharing data, privacy issues (real and perceived), and the lack of financial models to incentivize storage, sharing, and integration. Structured (eg, laboratory tests, demographics, diagnosis) and unstructured (eg, doctor notes, pathology reports, radiology reports) data in EHRs are not currently standardized, nor can current natural language-processing methods routinely and accurately extract handwritten notes within EHR systems to use these data more efficiently.[6]

Managing and using big data may necessitate certain requirements unique to these data sets, including support for all data formats; data mobility; easy access through industry-standard application programming interfaces;

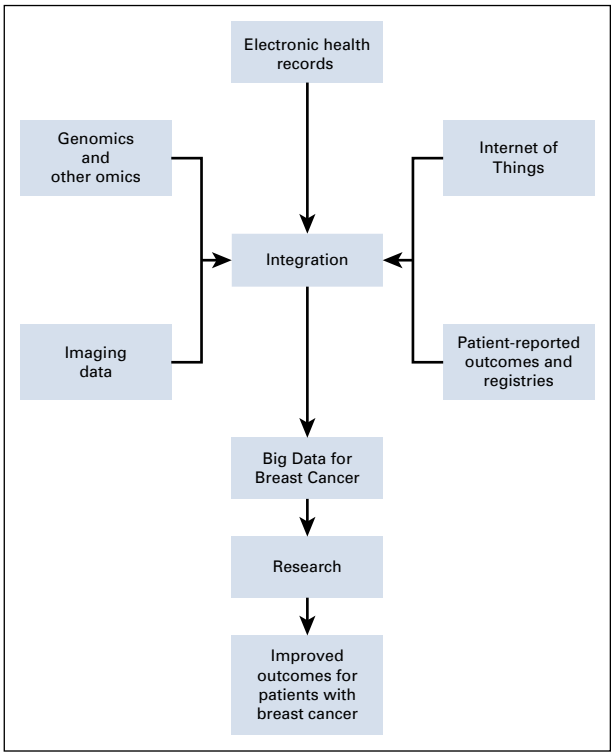

FIG 2. The amount of data related to breast cancer has exploded. The integration of various types of data (patient-reported outcomes; imaging data; electronic health record data; genomics and other omics; and so on) can fuel scientific discoveries and lead to improved outcomes for patients with breast cancer.

multiple processing engines; and a strong, unified security model. To meet these prerequisites in breast cancer, a field with extensive research and clinical data, investments must be made. This includes the technology to support development of these systems and applications and harness the power of these enhanced data sets in big data projects. Although data storage remains a challenge, many companies are creating scalable solutions for data storage, including health data.

Technologic innovation brings opportunities to accelerate research discovery and improve patient care. Precision medicine can be achieved by using big data technologies to combine different types and sources of data to identify patterns, determine optimal treatments for individuals, and improve their outcomes.[8] Imaging is another technologic opportunity where big data can be leveraged in breast cancer. Technologic advancements are transforming how physicians interact with clinical images. Approximately 75 million mammograms are performed annually worldwide, currently requiring a large amount of personnel time to read and analyze the results of each test.

Emerging advances in artificial intelligence (AI) methods (eg, convolutional neural networks and deep learning) are being used in mammography to distinguish patients with benign, malignant, and negative disease,[9] and objectively assess breast density.[10] Similar advances are taking place in magnetic resonance imaging, where machine learning is being used to predict patients' response to therapy and outcomes.[11] Big data imaging applications can be used to automate image reading and generate reports without human interaction, allowing radiologists to only review images not easily interpreted by the algorithms. Images can also be computer-enhanced, highlighting features to guide radiologists' expert interpretation.[12] Radiologists can then focus on reading images with complex features, reviewing and designing personalized treatment plans, and providing overall quality control to constantly deliver optimal patient care.[13] Preliminary results suggest that the best outcome of incorporating machine learning in radiology occurs when radiologists are part of the decision process in reviewing and approving the output of the algorithms.[14]

RESEARCH APPLICATION

Much hope has been placed in using real-world data (RWD; data collected outside clinical trials) or real-world evidence (RWE, RWD plus analytics) to complement knowledge gathered from clinical trials.[15] Successfully using RWD and RWE requires numerous factors: aggregated, high-quality, complete, longitudinal data sets; reproducibility and provenance; patient-level data linkage; end points and outcomes; study objectives and analysis plans; and careful cohort selection. Social, ethical, and legal challenges exist for using RWD. Although these data are often not standardized, natural language processing and other advanced technologies such as AI can help simplify extraction of unstructured data.[16]

Another challenge is that big data are less well curated than other classic data sets (eg, prospective clinical trials and epidemiologic registries). This causes garbage-in–garbage-out concerns. Perhaps the biggest barrier in big data research application is that basic laboratory and clinical translational researchers often have no expertise in the multidimensional analytics and visualization tools that big data requires. Research funding organizations (public and private) often do not provide support to cover costs of data annotation, curation, and sharing, and they rarely support training programs focused on big data.

Applying big data to research applications offers enormous possibilities, with the potential to solve questions currently unanswerable in traditional research laboratories. Big data approaches are particularly powerful with cross-platform analyses, such as combining EHRs or clinical trial outcomes data with omics data sets. For example, The Cancer Genome Atlas collects data (omics, RNA and DNA sequences, expression information, and clinical metadata) from more than 10,000 patients with cancer and 33 different types of cancer.[17] With various tools to interrogate The Cancer Genome Atlas database, drivers of endocrine therapy resistance in breast cancer were identified.[18]

Machine learning was successfully used on those data to identify miRNA biomarkers in breast cancer.[19]

PROs, RWD, and RWE represent other areas where big data can be generated and harnessed for research applications.[15] PROs include data collected directly from patients (eg, quality of life and functional status) that are not interpreted by physicians or others.[20] Because only approximately 3% of patients with cancer participate in clinical trials, big data approaches to PROs and RWD/RWE may represent realistic ways of integrating information about under-represented populations and finding solutions to previously intractable clinical problems. Using EHR or IoT data sets, in contrast to clinical trials, allows interrogation of extremely large numbers of patients and generation of enormous volumes of RWD. Passing of the 21st Century Cures Act[21] and the Prescription Drug User Fee Act Reauthorization[22] has resulted in regulatory guidance on how to interrogate this information. Flatiron Health has created a clinicogenomic database containing real-world, longitudinal, patient-level clinical EHR data from cancer clinics. Their goal is to use RWD to answer clinical questions. These data are linked to deep, next-generation sequence profiling across hundreds of cancer-related genes for each patient's tumor, as assessed by Foundation Medicine (Cambridge, MA).[23] Pharmaceutical companies and the US Food and Drug Administration are also active in the field and understand the value of RWD/RWE in developing and regulating medical products.[24,25]

Komen's BD4BC meetings offered valuable insights into the great and unmet need to advance training in and improve access to big data. Breast cancer researchers without a data science focus could be trained in the field of data science by participating in workshops and conferences to learn new methods, skills, and techniques to advance their research projects using big data. Effective use of big data to improve breast cancer care can also be facilitated by bringing together data set owners, both nonprofit and for-profit, with data scientists, with the goal of creating access to breast cancer data sets for research. Komen and similar organizations are well positioned to meet this demand. For example, new award programs could be established to support data sharing, multidisciplinary research projects, and cross training of investigators.

PRIVACY AND DATA SHARING

The roles of laws and regulations, institutional constraints, and patients themselves in how data are both protected and shared need to be considered.[7] Many patients are motivated to share their data,[26] as shown by patients' willingness to actively participate in data sharing in research projects such as the Metastatic Breast Cancer Project. Several thousand patients with metastatic breast cancer have registered to share their data for research.[27] A similar participation response was recently obtained by the National Institutes of Health (NIH) with their All of Us study.[28] Patients' desire for new knowledge about their disease often outweighs their privacy concerns. Still, many patients have questions about what control they have over their data, electronic accessibility,[26] privacy, and data security, and why data sharing is important.

Government laws and regulations make medical data one of the few remaining bastions of privacy. The Health Insurance Portability and Accountability Act of 1996 and the Office of Human Research Protection limit access to patient records and tissues. This makes research challenging because much of the medical data are housed in data silos and unavailable for clinical or research use. In addition to the Health Insurance Portability and Accountability Act and Office of Human Research Protection, siloed data have additional barriers, including business or proprietary interests, privacy concerns, transaction costs associated with data infrastructure and data sharing, and the nature of local legacy systems.

Although data sharing has become more common, it is often inefficient because data must be standardized, de-identified, and possibly shared via an institutional review board–approved study with appropriate participant consent. This presents another challenge because most patients only agree to one study at a time, meaning re-consent is needed for each use.[29] The growing number of companies that view patient data as a commercial asset is particularly concerning. What a patient originally agreed to share with his or her initial consent may be unknown. Whether data can be shared for additional investigations and commercial purposes is unknown.[6] Other challenges to data sharing include interface glitches, moving data between formats, and questions of whether the data depositor is legally liable if public data are misinterpreted by another user.[7] Some companies do not want their data placed in a centralized, accessible location because they want to retain control of them, retain exclusive rights to their analyses, and/or monetize the data and their usage. Overall, little incentive exists to share or make data readily and easily accessible.

Patient advocacy organizations such as Komen can help address some of the problems associated with data sharing, starting with educating and advocating for patient needs and concerns regarding privacy and data sharing. In turn, the government can resolve issues about multiagency informed consent and establish privacy rules and structures to allow data sharing while protecting patients. Some research funding organizations (NIH, Komen, the Bill and Melinda Gates Foundation, and so on) require funded researchers to share their data.[7,30] Key agencies (NIH, National Science Foundation, European Research Council, and Research Councils UK) promote the concept of open data, where data are released to public databases.[1] Other examples, such as NIH's Gene Expression Omnibus and cBioPortal, intend to place data in public databases so researchers can integrate them and learn from them. The various layers of data should be present in platforms that allow integration and enable

a deep dive into what is happening to individual patients. A link back to the individual (while maintaining security and privacy) for additional data collection is preferable. Financial support for infrastructure, curation, de-identification, sustainability, and appropriate guidelines is necessary to implement this requirement.[1]

Another highlight from Komen's BD4BC meetings is that motivation for sharing data may increase if patients learn the value of sharing information for big data-driven research and if the study results are shared with patients. Komen is committed to empowering the patient community with information and tools to make data sharing understandable and easy. Training to empower patient advocates to participate in the data science research process should be offered. This highlights the important role patient advocates can play in demanding that data be shared, such as pressuring different stakeholders to work together toward a common goal, urging sustainability of good ideas, and requesting that study results be shared with patients. Including patient advocates early in discussions can lead to better, faster, and more accepted results by the public. This is critical for moving big data efforts forward. Better knowledge about patients' decision making regarding data sharing will lead to more effective distribution of information to help them make key decisions about their health and care.

PATIENT CARE APPLICATION

Many emerging technologies, such as the previously mentioned omics platforms and wearable devices, are beginning to generate actionable, individual-specific data that can be used to improve patient care. Historically, the tumor of a patient with breast cancer might be evaluated at the protein level for hormone receptors, human epidermal growth factor receptor 2, and Ki67 and at the transcriptome level with a multigene assay such as Oncotoype Dx (Genomic Health, Redwood City, CA), MammaPrint (Agendia, Irvine, CA), or others.[31] Today, that same tumor can undergo whole exome, genome, or deep sequencing. Investigations of the microbiome[32] and liquid biopsies[33] are other emerging data sources. Such data can provide guidance on the likelihood a given patient with breast cancer will benefit from chemotherapy.[34] This type of testing has also stratified more than 15 subtypes of breast cancer, allowing omics data to inform prognoses and treatments.[3,35,36]

Cases showing the actual value of big data in patient care are currently lacking. A few anecdotal examples of how big data are being integrated in the patient care workflow exist. For example, one of us (M.L.) created an analytic dashboard to leverage patient data and show patients' care status in near–real time. It was used to evaluate the institution's policy to delay breast biopsies by 7 days for women taking any type of anticoagulants. The dashboard ultimately led to changing the policy and to women undergoing biopsy sooner.

For millions of patients with breast cancer, EHRs represent an important, comprehensive resource for patient care. Because patients continuously receive care, EHRs allow longitudinal tracking of long-term outcomes (eg, time receiving therapy and safety events). EHR-based treatment plans are also being used to reduce medication errors compared with prior paper-based approaches, increase standardization, and allow retrieval of data for quality measures within institutions.

Challenges remain with EHR systems. A primary goal of big data should be to optimize doctors' ability to care for patients. With the integration of omics analyses in EHRs to drive clinical decision, EHR systems have become large enterprises, requiring large infrastructure and computational power many clinics cannot afford. Current EHRs frequently decrease clinical efficiency, demoralize physicians by turning them into data entry specialists, and decrease time of direct patient-physician interaction. Physicians feel EHR systems exist primarily to aid hospital billing, rather than facilitate patient care. One of us (G.S.) highlighted what was at stake in improving how physicians interact with EHR systems to optimize patient care: "Save a doctor's time, save a patient's life."

An identified opportunity for EHR systems is their ability to revolutionize how quality of care is assessed. Quality measures (eg, ASCO's Quality Oncology Practice Initiative metrics) should flow seamlessly from EHRs, and clinical decision support should become a standard aspect of EHRs. Including genomic and other omics data directly in EHRs should promote high-level clinical decision support and improved access to clinical trials to realize the full potential of precision medicine. The expected result is higher-quality, more cost-effective treatment that is precise and targeted to the patient.

HEALTH EQUITY

Although projects using genomic data to inform treatment decisions show promising results in breast cancer, the use of big data to reduce health disparities in breast cancer care is limited.[37] A recent example is the ACCURE (Accountability for Cancer Care Through Undoing Racism and Equity) quality improvement trial, which focused on the racial disparity in completing treatment of curable breast cancer. This disparity contributes to worse survival among African Americans. This trial used an EHR system to create a real-time registry of patients that alerted the health care team when participants missed appointments or had an unmet care milestone. This system was paired with patient navigation and clinical feedback. The intervention improved treatment completion and helped reduce racial disparities in treatment of these patients.[38] ACCURE exemplifies the power of harnessing big data to address health disparities in breast cancer care. More of these types of projects are needed.

In some cases, big data itself could be a source of disparities. The resources, knowledge, and infrastructure needed to use

big data may not be available to all care providers and patients. The patient populations at health systems with sophisticated EHRs, data sharing systems, and capabilities to collect and store biospecimens for later analysis are likely not representative of the true diversity of patients across all settings.[37] Thus, any advancement resulting from big data should be designed and implemented for the maximum number of individuals who can benefit from it. Minority populations should be adequately represented in data sets so that safe and effective treatment strategies for all patients can be drawn from the results of studies using them.

Increasing efforts to better classify breast cancer are driving wide use of molecularly based precision medicine in oncology practice. Data have always been the backbone of epidemiology and population health studies to investigate breast cancer incidences and identify interventions that affect outcomes. Access to big data can supercharge these types of data-fueled studies by providing an ever-growing number of data points and characteristics about everyone within a population.[37] AI may be able to identify patterns within subpopulations that predict the risk of occurrences, recurrences, and generally worse outcomes, as well as optimal, tailored treatment plans. These studies may allow more efficient access to care, better treatment adherence by patients within the continuum of care, and more precise identification and management of at-risk populations. Big data may truly facilitate personalized medicine, regardless of race, ethnicity, or other characteristics.

DISCUSSION

With the growing availability of PROs and RWD, AI will likely affect clinical practice and clinical trial design in the near term, and increase our understanding of patients' response to therapy.

Yet, most of the general public and many patients are unaware of or have concerns about big data and its application to cancer research and treatment. Developing and implementing educational resources (eg, fact sheets, Web portal on big data, patient advocate training) will be critical to making data sharing understandable and easy. Patients want to control the use of their data.[4] An educated, engaged advocacy community is crucial for BD4BC to succeed. Komen will start by developing an online knowledge portal designed as a hub of information for visitors to advance their knowledge about big data.

Meeting participants repeatedly mentioned the lack of both data scientists working in breast cancer and laboratory and clinical researchers who are fluent in big data analytics. Komen and similar organizations can support researcher education by providing funding opportunities to train breast cancer researchers in big data and attract data scientists to apply big data to solve remaining challenges in breast cancer.[39] A central directory of existing breast cancer data sets available to big data scientists to use in breast cancer research is essential.

In addition, support must exist for projects that will accelerate the technologic advancement and innovative thinking necessary to discover novel targets for precision medicine and provide earlier detection that will affect

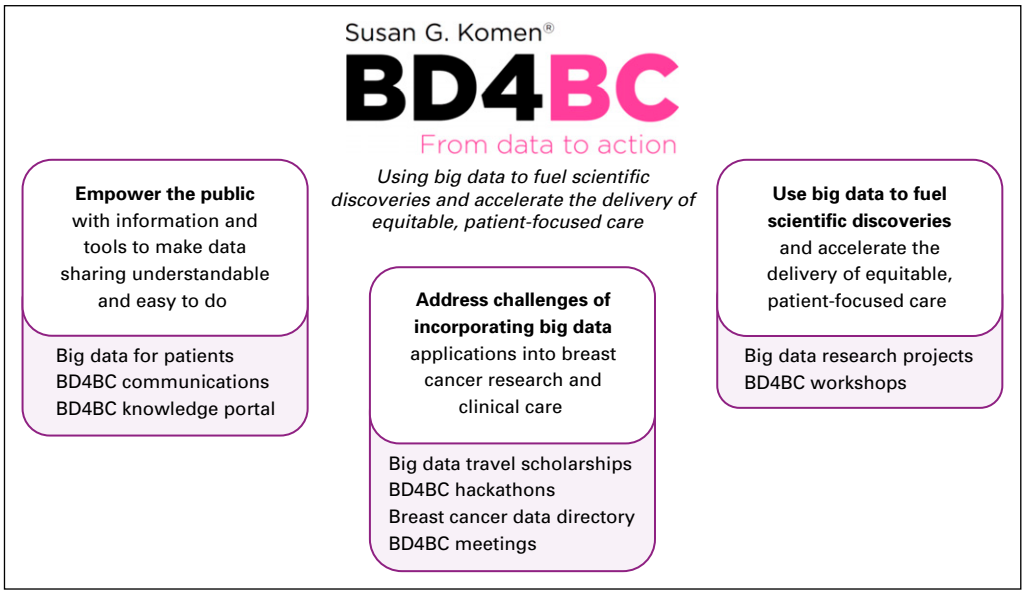

FIG 3. Komen's Big Data for Breast Cancer (BD4BC) initiative. The initiative is aimed at using big data to fuel scientific discoveries and accelerate the delivery of equitable, patient-focused care. Komen is following a three-pronged approach (white boxes) in developing several programs (pink boxes) that will specifically address some of the challenges identified during Komen's BD4BC meetings and detailed in this article. Bold text indicates the main action items of this initiative.

patients with breast cancer, including those living with metastatic breast cancer. Organizations such as Komen can leverage their grant-making capabilities to identify and support big data research resources and projects that put the patient at the center of innovation to inform and accelerate the pace of breast cancer research and improve patient care.

Funding is needed to support data science projects and infrastructure for aggregation, visualization, and modeling of patient- and laboratory-derived big data. Review criteria, eligibility requirements, and expenses allowed under such funding must change to accommodate the specific needs of data science projects. Public advocacy is needed so initiatives can leverage big data to support adherence to treatment and participation in clinical trials, and enforce safety, security, data standards, accessibility, and sharing requirements. Protection of minorities and underserved populations in the big data revolution is needed to ensure these groups are included in the progress big data will make toward better breast cancer outcomes.

In conclusion, many opportunities were identified at Komen's BD4BC meetings to harness big data to benefit patients with breast cancer. Now is the time to move forward. Komen is working with partners to design BD4BC initiatives to improve outcomes for patients with breast cancer (Fig 3). Strategies include educating the public to make data sharing understandable and easy, addressing challenges of incorporating big data applications into breast cancer research and clinical care, and funding data science projects. Komen will continue to advocate for putting the patient at the center of innovation and support efforts using big data to fuel scientific discoveries and accelerate the delivery of improved, equitable, and patient-focused care. Komen invites other organizations to join in realizing this big data revolution.

AFFILIATIONS

[1]Susan G. Komen, Dallas, TX
[2]University of Kansas Cancer Center, Kansas City, KS
[3]Rush University Medical Center, Chicago IL
[4]Vanderbilt University, Nashville, TN
[5]Stanford University School of Medicine, Stanford, CA

CORRESPONDING AUTHOR

Jerome Jourquin, PhD, MS, Susan G. Komen, 5005 LBJ Freeway, Ste 526, Dallas, TX 75244; Twitter: @SusanGKomen; e-mail: JJourquin@komen.org.

EQUAL CONTRIBUTION

J.J. and S.B.R. are equal co-first authors of this work.

ACKNOWLEDGMENT

The authors thank the Robertson Foundation for its support of Komen's first two BD4BC meetings. Komen acknowledges the critical role Nancy G. Brinker played in creating this vision and bringing Komen's BD4BC initiative to fruition. We thank all members of the Planning Committees for Komen's three BD4BC meetings for all their efforts in guiding agendas and recommending participants and speakers. We also thank the many speakers, panelists, advocates, and participants who gave their time, expertise, and energy to inform and guide this work. The authors thank Kristine De La Torre, PhD, and Elizabeth Gordon, PhD, MPH, for medical writing assistance.

AUTHOR CONTRIBUTIONS

Conception and design: Jerome Jourquin, Stephanie Birkey Reffey, Cheryl Jernigan, Mia Levy, Glendon Zinser, Kimberly Sabelko, George Sledge Jr

Administrative support: Jerome Jourquin, Stephanie Birkey Reffey, Glendon Zinser

Collection and assembly of data: Jerome Jourquin, Stephanie Birkey Reffey

Data analysis and interpretation: Jerome Jourquin, Stephanie Birkey Reffey, Glendon Zinser, Kimberly Sabelko, Jennifer Pietenpol, George Sledge Jr

Manuscript writing: All authors

Final approval of manuscript: All authors

Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/po/author-center.

Stephanie Birkey Reffey
Travel, Accommodations, Expenses: Genentech

Mia Levy
Employment: SeqTech Diagnostics (I)
Leadership: Personalis
Stock and Other Ownership Interests: Personalis, GenomOncology
Honoraria: Roche
Consulting or Advisory Role: Personalis, GenomOncology, Roche
Research Funding: GenomOncology
Patents, Royalties, Other Intellectual Property: Royalties from GenomOncology for licensing of MyCancerGenome content
Travel, Accommodations, Expenses: Roche

Jennifer Pietenpol
Stock and Other Ownership Interests: Bluebird Bio, Johnson & Johnson, Roche, Gilead Sciences (I), Quest Diagnostics, Illumina, Kite Pharma, Novartis (I)
Research Funding: Incyte Corporation (Inst)
Patents, Royalties, Other Intellectual Property: Jennifer Pietenpol is an inventor (PCT/US2012/065724) of intellectual property (TNBCtype) licensed by Insight Genetics
Other Relationship: Susan G. Komen

George Sledge Jr
Leadership: Syndax, Tessa Therapeutics
Stock and Other Ownership Interests: Syndax, Tessa Therapeutics
Consulting or Advisory Role: Radius Health, Taiho Pharmaceutical, Symphogen, Synaffix, Syndax, Verseau Therapeutics
Research Funding: Genentech (Inst), Pfizer (Inst)
Travel, Accommodations, Expenses: Radius Health, Verseau Therapeutics, Tessa Therapeutics

No other potential conflicts of interest were reported.

REFERENCES

1. Leonelli S: Why the current insistence on open access to scientific data? Big data, knowledge production, and the political economy of contemporary biology. Bull Sci Technol Soc 33:6-11, 2013
2. Wu PY, Cheng CW, Kaddi CD, et al: -Omic and electronic health record big data analytics for precision medicine. IEEE Trans Biomed Eng 64:263-273, 2017
3. Flores M, Glusman G, Brogaard K, et al: P4 medicine: How systems medicine will transform the healthcare sector and society. Per Med 10:565-576, 2013
4. Chung AE, Jensen RE, Basch EM: Leveraging emerging technologies and the "Internet of Things" to improve the quality of cancer care. J Oncol Pract 12:863-866, 2016
5. Odgers DJ, Dumontier M: Mining electronic health records using linked data. AMIA Jt Summits Transl Sci Proc 2015:217-221, 2015
6. Maggi N, Gazzarata R, Ruggiero C, et al: Cancer precision medicine today: Towards omic information in health care systems. Tumori 105:38-46, 2019
7. Rosenbaum L: Bridging the data-sharing divide—seeing the devil in the details, not the other camp. N Engl J Med 376:2201-2203, 2017
8. Yarchoan M, Hopkins A, Jaffee EM: Tumor mutational burden and response rate to PD-1 inhibition. N Engl J Med 377:2500-2501, 2017
9. Aboutalib SS, Mohamed AA, Berg WA, et al: Deep learning to distinguish recalled but benign mammography images in breast cancer screening. Clin Cancer Res 24:5902-5909, 2018
10. Lehman CD, Yala A, Schuster T, et al: Mammographic breast density assessment using deep learning: Clinical implementation. Radiology 290:52-58, 2018
11. Tahmassebi A, Wengert GJ, Helbich TH, et al: Impact of machine learning with multiparametric magnetic resonance imaging of the breast for early prediction of response to neoadjuvant chemotherapy and survival outcomes in breast cancer patients. Invest Radiol 54:110-117, 2019
12. European Society of Radiology (ESR): What the radiologist should know about artificial intelligence—an ESR white paper. Insights Imaging 10:44, 2019
13. Kansagra AP, Yu JP, Chatterjee AR, et al: Big data and the future of radiology informatics. Acad Radiol 23:30-42, 2016
14. Barzilay R: 30th Anniversary AACR Special Conference Convergence: Artificial intelligence, big data, and prediction in cancer. Presented at the American Association for Cancer Research, Newport, RI, October 14-17, 2018
15. Sherman RE, Anderson SA, Dal Pan GJ, et al: Real-world evidence—what is it and what can it tell us? N Engl J Med 375:2293-2297, 2016
16. Usui M, Aramaki E, Iwao T, et al: Extraction and standardization of patient complaints from electronic medication histories for pharmacovigilance: Natural language processing analysis in Japanese. JMIR Med Inform 6:e11021, 2018
17. Sun Q, Li M, Wang X: The Cancer Omics Atlas: An integrative resource for cancer omics annotations. BMC Med Genomics 11:63, 2018 [Erratum: BMC Med Genomics 11:74, 2018]
18. Anurag M, Punturi N, Hoog J, et al: Comprehensive profiling of DNA repair defects in breast cancer identifies a novel class of endocrine therapy resistance drivers. Clin Cancer Res 24:4887-4899, 2018
19. Sherafatian M: Tree-based machine learning algorithms identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping. Gene 677:111-118, 2018
20. Weldring T, Smith SM: Patient-reported outcomes (PROs) and patient-reported outcome measures (PROMs). Health Serv Insights 6:61-68, 2013
21. Lye CT, Forman HP, Daniel JG, et al: The 21st Century Cures Act and electronic health records one year later: Will patients see the benefits? J Am Med Inform Assoc 25:1218-1220, 2018
22. Schneeweiss S, Glynn RJ: Real-world data analytics fit for regulatory decision-making. Am J Law Med 44:197-217, 2018
23. Agarwala V, Khozin S, Singal G, et al: Real-world evidence in support of precision medicine: Clinico-genomic cancer data as a case study. Health Aff (Millwood) 37:765-772, 2018
24. Singh G, Schulthess D, Hughes N, et al: Real world big data for clinical research and drug development. Drug Discov Today 23:652-660, 2018
25. Khozin S, Blumenthal GM, Pazdur R: Real-world data for clinical evidence generation in oncology. J Natl Cancer Inst 109: 2017
26. Haug CJ: Whose data are they anyway? Can a patient perspective advance the data-sharing debate? N Engl J Med 376:2203-2205, 2017
27. Wagle N: 30th Anniversary AACR Special Conference convergence: Artificial intelligence, big data, and prediction in cancer, partnering with patients to advance cancer research. Presented at the American Association for Cancer Research, Newport, RI, October 14-17, 2018
28. Denny J: AACR modernizing population sciences in the digital age, early progress on the All of Us research program. Presented at the American Association for Cancer Research, San Diego, CA, February 19-22, 2019
29. Ludman EJ, Fullerton SM, Spangler L, et al: Glad you asked: Participants' opinions of re-consent for dbGap data submission. J Empir Res Hum Res Ethics 5:9-16, 2010
30. Wykstra S: Funder data-sharing policies: Overview and recommendations. https://www.healthra.org/wp-content/uploads/2018/08/RWJF-final-report-for-Figshare.pdf
31. Vieira AF, Schmitt F: An update on breast cancer multigene prognostic tests-emergent clinical biomarkers. Front Med (Lausanne) 5:248, 2018
32. Fernández MF, Reina-Pérez I, Astorga JM, et al: Breast cancer and its relationship with the microbiota. Int J Environ Res Public Health 15:E1747, 2018
33. Chu D, Park BH: Liquid biopsy: Unlocking the potentials of cell-free DNA. Virchows Arch 471:147-154, 2017
34. Sparano JA, Gray RJ, Makower DF, et al: Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. N Engl J Med 379:111-121, 2018
35. Lehmann BD, Bauer JA, Chen X, et al: Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Invest 121:2750-2767, 2011
36. Lehmann BD, Jovanović B, Chen X, et al: Refinement of triple-negative breast cancer molecular subtypes: Implications for neoadjuvant chemotherapy selection. PLoS One 11:e0157368, 2016
37. Reeder-Hayes KE, Troester MA, Meyer AM: Reducing racial disparities in breast cancer care: The role of 'big data.' Oncology (Williston Park) 31:756-762, 2017
38. Cykert S, Eng E, Manning MA, et al: A multi-faceted intervention aimed at black-white disparities in the treatment of early stage cancers: The ACCURE Pragmatic Quality Improvement trial. J Natl Med Assoc 10.1016/j.jnma.2019.03.001. [Epub ahead of print on March 28, 2019]
39. Ibnouhsein I, Jankowski S, Neuberger K, et al: The big data revolution for breast cancer patients. Eur J Breast Health 14(2):61-62, 2018