CHAPTER 13

The Promise and Potential Perils of Big Data for Advancing Symptom Management Research in Populations at Risk for Health Disparities

Suzanne Bakken and Nancy Reame

ABSTRACT

Symptom management research is a core area of nursing science and one of the priorities for the National Institute of Nursing Research, which specifically focuses on understanding the biological and behavioral aspects of symptoms such as pain and fatigue, with the goal of developing new knowledge and new strategies for improving patient health and quality of life. The types and volume of data related to the symptom experience, symptom management strategies, and outcomes are increasingly accessible for research. Traditional data streams are now complemented by consumer-generated (i.e., quantified self) and “omic” data streams. Thus, the data available for symptom science can be considered big data. The purposes of this chapter are to (a) briefly summarize the current drivers for the use of big data in research; (b) describe the promise of big data and associated data science methods for advancing symptom management research;

(c) explicate the potential perils of big data and data science from the perspective of the ethical principles of autonomy, beneficence, and justice; and (d) illustrate strategies for balancing the promise and the perils of big data through a case study of a community at high risk for health disparities. Big data and associated data science methods offer the promise of multidimensional data sources and new methods to address significant research gaps in symptom management. If nurse scientists wish to apply big data and data science methods to advance symptom management research and promote health equity, they must carefully consider both the promise and perils.

INTRODUCTION

Symptom management research is a core area of nursing science and one of the priorities for the National Institute of Nursing Research, which specifically focuses on understanding the biological and behavioral aspects of symptoms such as pain and fatigue, with the goal of developing new knowledge and new strategies for improving patient health and quality of life (National Institute of Nursing Research, 2015). The types and volume of data related to the symptom experience, symptom management strategies, and outcomes are increasingly accessible for research. Traditional data streams are now complemented by consumer-generated (i.e., quantified self) and “omic” data streams (Figure 13.1) (Dodd et al., 2001; Swan, 2013). Thus, the data available for symptom science can be considered big data. Nurse scientists have integrated genomics into their programs of research in symptom management (Cashion et al., 2013; Merriman et al., 2014; Reddy et al., 2014), but less research has focused on the other big data sources such as social media to capture the symptom experience and symptom management strategies (Yoon, Elhadad, & Bakken, 2013). Authors are increasingly noting the vital role of quantified self (phenome) and environmental (envirome) big data for research as well as personalized medicine (Dudley, Listgarten, Stegle, Brenner, & Parts, 2015; Hansen, Miron-Shatz, Lau, & Paton, 2014; Swan, 2013).

Big data are typically conceptualized in terms of volume, variety, and velocity (e.g., sensor or streaming data) and more recently, veracity, which refers to the level of uncertainty associated with the collection of data sources (IBM, 2015). These multiple attributes of big data reflect an additional complexity beyond that of simple volume (e.g., Medicare claims data). Some authors also consider a fifth “V” in data science attributes—value, an attribute that points to the necessity of the rigorous methods of data science to extract knowledge from data so that it is of value (Marr, 2015). The National Consortium for Data Science has emphasized value in its definition of data science as “the systematic study of the

UCSF Symptom Management Model, 2001

FIGURE 13.1 Data streams for symptom management research.

Note. SNP = single nucleotide polymorphism.

*consumer available, Swan (2013).

organization and use of digital data in order to accelerate discovery, improve critical decision-making processes, and enable a data-driven economy” (Ahalt et al., 2014, p. iii). Data science draws upon theories and techniques from multiple fields such as mathematics, statistics, and information technology and includes methods such as probability models, machine learning, statistical learning, computer programming, pattern recognition and learning, visualization, predictive analytics, uncertainty modeling, data warehousing, and high-performance computing (Wikipedia, 2015).

The purposes of this chapter are to (a) briefly summarize the current drivers for the use of big data in research; (b) describe the promise of big data and associated data science methods for advancing symptom management research; (c) explicate the potential perils of big data and data science from the perspective of the ethical principles of autonomy, beneficence, and justice; and (d) illustrate strategies for balancing the promise and the perils of big data through a case study of a community at high risk for health disparities.

WHAT ARE THE CURRENT DRIVERS FOR THE USE OF BIG DATA IN RESEARCH?

There are several important drivers for the use of big data in research that are relevant to advancing the science of symptom management. First, electronic health-related data have increased in volume and variety. Consumer-generated data from mobile apps, wearables, medical devices, and sensors along with multiple sources of environmental and “omic” data complement traditional electronic sources of research data such as surveys, clinical data warehouses, and transaction data. Second, the number of research networks has expanded. For example, the development of the National Patient-Centered Clinical Research Network by the Patient-Centered Outcomes Research Institute (PCORI) has increased the availability of data from a variety of data sources through its Clinical Data Research Networks and Patient-Powered Research Networks (PCORI, 2015). Also, the Clinical Translational Science Award Accrual to Clinical Trials Network was launched recently to develop a nationwide network of sites that share electronic health record data (National Center for Advancing Translational Sciences, 2015). Third, initiatives related to precision medicine and biobanking are growing, including the recent launch of President Obama’s Precision Medicine Initiative (Collins & Varmus, 2015). Fourth, through the Big Data to Knowledge initiative (National Institutes of Health [NIH], 2015), the NIH has acknowledged the role of big data to advance scientific knowledge and launched a set of funding opportunities designed to develop resources and increase the competency of the biomedical workforce in data science in three core areas: computer science, biostatistics, and biomedical science.

WHAT IS THE PROMISE OF BIG DATA FOR ADVANCING SYMPTOM SCIENCE IN POPULATIONS AT RISK FOR HEALTH DISPARITIES?

In terms of the promise of big data for advancing symptom science in populations at risk for health disparities, research studies suggest that there are differences—which have been attributed to race and ethnicity—in symptom experiences, symptom management strategies, and symptom outcomes. In terms of the symptom experience, illustrative studies report

- Higher rates of pain among multiracial individuals with lung and colorectal cancer (Martinez, Snyder, Malin, & Dy, 2014)
- Increased severity of pain in Black patients with lung and colorectal cancer (Martinez et al., 2014)
- Lower rates of 14 physical symptoms among Asian Americans (Bauer, Chen, & Alegria, 2012)

- Decreased depression symptom severity among African Americans exposed to trauma (Ghafoori, Barragan, Tohidian, & Palinkas, 2012)
- Risk for moderate lower urinary tract symptoms higher in Hispanic and Black men and lower in Asian men (Van Den Eeden et al., 2012)
- Black women younger than 50 years more likely to report frequent and intense prodromal symptoms of myocardial infarction (McSweeney et al., 2010)

Studies have also reported racial differences in symptom strategies. For example, there are lower rates of recognition of late-life depression and treatment in Blacks (Pickett, Greenberg, Bazelais, & Bruce, 2014). In terms of differences in symptom outcomes, one example is that the time to reach functional decline in multiple sclerosis is shorter in Black Caribbeans than British Whites (Koffman et al., 2013).

A few recent studies are briefly described to support the potential of a variety of big data sources working in parallel to increase the understanding of the symptom experience. In a study that examined genes, environment, and depression symptoms in adolescents, female 5-HTTLPR genotype carriers had a marginally significant protective effect for depressive symptoms and no significant influences of environment (Uddin et al., 2011). In contrast, males residing in neighborhoods with relatively poor building conditions had higher depressive symptoms but no differences related to genotype. There were no significant gene–environment interactions.

Important distinctions between genotype and phenotype symptom expression, especially in underserved minorities, can be elucidated with big data approaches. One area of special interest for symptom research relates to racial/ethnic differences in drug pharmacokinetics, as clearance rates are known to affect treatment efficacy as well as risk for adverse effects. For example, in a study of the drug olanzapine, used in the treatment of acute psychosis, genotype rather than race, as previously documented, was a significant predictor of clearance rate (Bigos et al., 2011). At standard antipsychotic doses, 50% of individuals with the high-clearance genotype have trough blood levels below the therapeutic range and more African Americans carry the allele. After accounting for CYP3A43 genotype, race was no longer a significant predictor of olanzapine clearance.

Through remote monitoring using a smart inhaler, Van Sickle et al. (Van Sickle, Magzamen, Truelove, & Morrison, 2013) combined data about the use of inhaled, short-acting bronchodilators and location and presented these data to 30 asthma patients electronically. This resulted in improved asthma control and a decline in day-to-day asthma symptoms. Participants' reports suggest that the mechanisms for this improvement included increased awareness and

understanding of asthma patterns, level of control, bronchodilator use (timing, location) and triggers, and improved preventive practices.

Studies such as these provide examples of how the use of big data can contribute to advancing the science of symptom management. However, the use of big data and data science methods for these purposes is in its infancy, and nurse scientists have much to offer regarding defining important questions, applying relevant theories, contributing to the design of new data science methods, and considering the ethical, legal, and social implications (Brennan & Bakken, 2015).

WHAT ARE THE POTENTIAL PERILS OF BIG DATA AND DATA SCIENCE FROM THE PERSPECTIVE OF THE ETHICAL PRINCIPLES OF AUTONOMY, BENEFICENCE, AND JUSTICE?

There exist a number of potential perils of big data related to advancing symptom management research in populations at risk for health disparities, most of which are also relevant to big data research beyond symptom science. These are briefly summarized in the following paragraphs using the three principles for the ethical conduct of research subjects as articulated in the historic Belmont Report: respect for persons (i.e., autonomy), beneficence, and justice (The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979).

The principle of respect for persons includes two separate moral requirements: the requirement to acknowledge autonomy and the requirement to protect those with diminished autonomy (The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979). The primary mechanism for protection of autonomy is informed consent that includes adequate information—understandable in lay terms—to make an informed decision about participation (Reame, 2013). Some big data sources have explicit opt-in or opt-out consent processes, and the use of protected health information for research has ethical and regulatory oversight from institutional review boards and the Health Insurance Accountability and Portability Act. However, users of social network sites and other quantified self technologies may not fully comprehend in what ways their data can be used and who may access them. Terms of service are usually lengthy and use of data for research or commercial purposes may not be highlighted. Consequently, although an individual may technically consent by clicking “I agree,” the consent does not meet typical research criteria. Kahn and colleagues (Kahn, Vayena, & Mastroianni, 2014) recommend that “approaches to informed consent must be reconceived

for research in the social-computing environment, taking advantage of the technologies available and developing creative solutions that will empower users who participate in research, yield better results, and foster greater trust” (p. 13678).

Beneficence involves optimizing benefits while minimizing risks. Methodological rigor is an ethical requirement, not just a scientific one, to ensure that scarce resources are used wisely and decisions are not made based on unsound findings. This requirement remains critical in big data, although the methods vary from traditional research designs (Vayena, Salathe, Madoff, & Brownstein, 2015). Thus, poor methodological rigor is a potential peril for big data research. For example, rigor is needed in terms of selection of appropriate data streams; understanding data provenance; data extraction techniques; calibration and recalibration of algorithms designed or selected for application to the data sources; handling of confounders and biases; and approaches for pattern recognition, uncertainty modeling, predictive analytics, and so forth. Other important perils relate to risks associated with loss of confidentiality and commodification of patient-/consumer-generated data. One way these risks occur is through presumption, a process by which digital content is both produced and consumed by individuals as websites are accessed, mobile health applications are used, messages are posted and responded to on social network sites, and tweeters are followed. Vayena et al. (2015) delineate the importance of context of use and differentiating between public health (e.g., digital disease detection for outbreaks) and commercial purposes. In the “digital patient opinion economy” (Lupton, 2013), commercial use may vary along a continuum from activities such as postmarket surveillance for drugs and medical devices to tailored advertisements for products or services. So although individuals may benefit from their presumption activities, they do not typically reap financial benefits from commodification of their data.

The principle of justice gives rise to moral requirements that there be fair procedures and equitable outcomes in the selection of research participants. In regard to big data, the investigator is typically using existing data sources that may result in biases against underrepresented groups. For example, Latinos are less likely than Whites or Blacks to use an app for health tracking (Fox & Duggan, 2013). Differences in those represented in clinical and quantified self data sources could result in a limited understanding of the range of symptom experiences and a lack of cultural context. Similarly, racial and ethnic minorities are less likely to participate in biobanks (Dang et al., 2014; Shaibi, Coletta, Vital, & Mandarino, 2013), thus hindering discoveries based on “omic” data sources that may be of particular relevance to race or ethnicity.

BALANCING THE PROMISE AND PERILS: THE WASHINGTON HEIGHTS INWOOD INFORMATICS INFRASTRUCTURE FOR COMPARATIVE EFFECTIVENESS RESEARCH PROJECT

The Washington Heights Inwood Informatics Infrastructure for Comparative Effectiveness Research (WICER) project is used as a case study to illustrate balancing the promise and the perils associated with big data. The overall goals of WICER were to understand the health of the Washington Heights Inwood (WaHI) community in order to improve it and to build an informatics infrastructure for comparative effectiveness research. Funded as part of the Agency for Healthcare Research and Quality's PROSPECT portfolio, WICER included creating a research data warehouse comprising data from multiple clinical sources (e.g., inpatient, outpatient, home care) about WaHI residents, building tools that supported exploration of the research data warehouse and conduct of comparative effectiveness research, and a large community survey of almost 6,000 residents (Sittig et al., 2012; Wilcox, Fort, & Bakken, 2014). For the last, bilingual community health workers gathered data about social determinants of health, symptoms (e.g., depression, fatigue, anxiety), health behaviors (e.g., medication adherence, physical activity), patient-reported outcomes (e.g., quality of life), and selected physical measures (i.e., blood pressures, body mass index; Lee, Boden-Albala, Larson, Wilcox, & Bakken, 2014) to complement the clinical data about WaHI residents in the research data warehouse. A subsample of several thousand participants was also asked to provide saliva and dried blood spot samples for biomarker (e.g., oxytocin), DNA, and exome analyses (Bakken, Suero-Tejeda, Yoon, Weng, & Reame, 2015). In terms of environment, geocoded food environment data are being integrated with survey data to enable analyses about the role of environment and behaviors on selected health outcomes, and the WICER survey and research data warehouse has been compared in terms of content with other behavioral surveys according to the County Health Rankings Framework (Yoon, Wilcox, & Bakken, 2013). Given these data sources, the WICER data meet the big data attributes of volume and variety, but not necessarily veracity and velocity. However, we continue to augment our data sources. In regard to the last attribute, we have applied data mining and other data science methods to examine one type of quantified self data, tweets related to physical activity, but have not limited the tweets to those within WaHI (Yoon, Elhadad, & Bakken, 2013). Despite these caveats, the WICER project offers some useful insights related to balancing the promise and perils of big data in patients at risk for health disparities.

The major research promise of WICER is the rich characterization of the WaHI community based on the variety of data. This was considered particularly important, given that the primarily Latino immigrant community experienced

significant levels of health disparities and that Latinos are often underrepresented in research (Glickman et al., 2008; Gwadz et al., 2010) and historically have low levels of participation in biobanks (Dang et al., 2014; Shaibi et al., 2013).

The key mechanism for acknowledgment of autonomy was a detailed consent form that was administered by a bilingual community health worker in the participant's language of choice (English or Spanish) during the community survey. Potential participants not only consented to the survey but also explicated their willingness for each of the following: (a) contact for potential participation in other research studies by Columbia investigators; (b) linkage of community survey data with other electronic clinical data (e.g., research data warehouse); (c) use of biospecimens by the WICER team for biomarker and genetic studies related to heart disease and stroke; (d) use of biospecimens by WICER for research in general; (e) use of biospecimens linked to WICER data by investigators outside the WICER team; and (f) long-term storage and use of biospecimens for health-related research (Bakken et al., 2015). Using this "stepped" approach, participants maintained autonomy over the use of their data. Consistent with other literature (Dash et al., 2014), WICER participants were more willing to share their data for specific reasons and with researchers they knew. In a sample of 2,290 WICER survey respondents, consent rates for biospecimens were highest for use by the WICER team for studies related to heart disease and stroke (58%) and lowest for long-term storage and use for health-related research by investigators outside the WICER team (53.2%; Bakken et al., 2015).

In regard to beneficence, because WICER was noninterventional, the potential risks associated with participation related to a loss of confidentiality and these risks were minimized by data storage on secure systems including tablet computers (Wilcox, Gallagher, Boden-Albala, & Bakken, 2012), as well as institutional regulations (e.g., institutional review board approval) and policies (e.g., data use agreement). At the same time, there were multiple strategies in place aimed at optimizing potential benefits from the perspective of various WaHI stakeholders. First, data were collected using reliable and valid measures including Patient-Reported Outcomes Measurement Information System (PROMIS) scales for depression, anxiety, fatigue, social role performance, and health-related quality of life (Bevans, Ross, & Cella, 2014; Cella et al., 2010). Second, taking advantage of the WICER's informatics infrastructure and the nature of the data and biospecimens collected, our dissemination plan included data sharing with researchers beyond the WICER study team. Third, aggregated community survey data were shared with WaHI community-based organizations and health-care organizations to inform program development and community outreach (Bakken, Suero-Tejeda, Bigger, Wilcox, & Boden-Albala, 2014).

Fourth, the WICER team conducted 20 participatory design sessions, primarily in Spanish, with a subsample of community survey participants to design and iteratively refine infographics for returning participants' data to them in a manner suitable for various levels of health literacy (Arcia et al., 2013; Arcia et al., 2015). It is hoped that such approaches can lay the foundation for greater community engagement and ultimately serve as cues to action for health behavior change. Sample infographics for PROMIS anxiety score and depression are shown in Figure 13.2. To further optimize the benefits of the project, resources were designed for sharing. These included the system that was developed to create the infographics (the Electronic Tailored Infographics for Community Education, Engagement, and Empowerment system; Velez, Bales, Arcia, & Bakken, 2014), the style guide (Arcia & Bakken, 2015) that provided the programming specifications for matching the type of data to infographic design, and infographic designs.

FIGURE 13.2 Example infographics for Patient-Reported Outcomes Measurement Information System anxiety score and Centers for Disease Control and Prevention 30-day depression score.

To support fair procedures and equitable outcomes in the selection of research participants (i.e., justice), WICER was explicitly designed to have multiple recruitment strategies (Lee et al., 2014). First, participants were recruited primarily from nonclinical settings including a community center and households. Second, the household sample started as a probability sample to recruit participants to reflect the demographics of WaHI. This was subsequently augmented by snowball sampling to take advantage of the connections among the community members. Although the final sample had more Latinos and more females than the community demographics as a whole, given the underrepresentation of Latinos in research and the high risk for health disparities of Latinos, this was considered to be justified. The enriched population of Latino participants ultimately shaped the infographic presentations of health data and allowed for the return of individual health information in a culturally competent manner to better promote behavior change.

CONCLUSION

Symptom management is a key area of research for nurse scientists. Big data and associated data science methods offer the promise of multidimensional data sources and new methods to address significant research gaps. The potential perils resulting from research using big data may be more problematic for those at high risk for health disparities. If nurse scientists wish to apply big data and data science methods to advance symptom management research and promote health equity, they must carefully consider the promise and perils.

ACKNOWLEDGMENTS

Manuscript preparation was supported by the Agency for Healthcare Research and Quality (R01HS022961) and the National Center for Advancing Translational Sciences (UL1TR000040).

REFERENCES

- Ahalt, S., Bizon, C., Evans, J., Erlich, Y., Ginsberg, G., Krishnamurthy, A., et al. (2014). *Data to discovery: Genomes to health*. A white paper from the National Consortium for Data Science. RENCI, University of North Carolina at Chapel Hill. Text. <http://dx.doi.org/10.7921/G03X84K4>
- Arcia, A., & Bakken, S. (2015). Style Guide: An interdisciplinary communication tool to support the process of generating tailored infographics from electronic health data using EnTICE3. *eGEMs (Generating Evidence & Methods to Improve Patient Outcomes)*, 3(1), 1120.
- Arcia, A., Bales, M. E., Brown, W., 3rd., Co, M. C., Jr., Gilmore, M., Lee, Y. J., et al. (2013). Method for the development of data visualizations for community members with varying levels of health literacy. *AMIA Annual Symposium Proceedings*, 2013, 51–60.

- Arcia, A., Suero-Tejeda, N., Bales, M. E., Merrill, J. A., Yoon, S., Woollen, J., et al. (2015). Sometimes more is more: Iterative participatory design of infographics for engagement of community members with varying levels of health literacy. *Journal of the American Medical Informatics Association*. <http://dx.doi.org/10.1093/jamia/ocv079>
- Bakken, S., Suero-Tejeda, N., Bigger, J. T., Wilcox, A., & Boden-Albala, B. (2014). Weaving a strong trust fabric through community-engaged research: Lessons from the WICER Project about digital infrastructure for the Learning Health System. *AMIA Joint Summits Translational Science Proceedings, eCollection 2014*, p. 126.
- Bakken, S., Suero-Tejeda, N., Yoon, S., Weng, C., & Reame, N. (2015). Sociodemographic factors and health literacy are related to consent for collection, use, and storage, of biospecimens in a Latino community sample. *AMIA Joint Summits Translational Science Proceedings, eCollection 2015*, p. 241.
- Bauer, A. M., Chen, C. N., & Alegria, M. (2012). Associations of physical symptoms with perceived need for and use of mental health services among Latino and Asian Americans. *Social Science in Medicine*, 75(6), 1128–1133. <http://dx.doi.org/10.1016/j.socscimed.2012.05.004>
- Bevans, M., Ross, A., & Cella, D. (2014). Patient-Reported Outcomes Measurement Information System (PROMIS): Efficient, standardized tools to measure self-reported health and quality of life. *Nursing Outlook*, 62(5), 339–345. <http://dx.doi.org/10.1016/j.outlook.2014.05.009>
- Bigos, K. L., Bies, R. R., Pollock, B. G., Lowy, J. J., Zhang, F., & Weinberger, D. R. (2011). Genetic variation in CYP3A43 explains racial difference in olanzapine clearance. *Molecular Psychiatry*, 16(6), 620–625. <http://dx.doi.org/10.1038/mp.2011.38>
- Brennan, P. F., & Bakken, S. (2015). Nursing needs big data and big data needs nursing. *Journal of Nursing Scholarship*, 47(5), 477–484. <http://dx.doi.org/10.1111/jnu.12159>
- Cashion, A., Stanfill, A., Thomas, F., Xu, L., Sutter, T., Eason, J., et al. (2013). Expression levels of obesity-related genes are associated with weight change in kidney transplant recipients. *PLoS One*, 8(3), e59962. <http://dx.doi.org/10.1371/journal.pone.0059962>
- Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., et al. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology*, 63(11), 1179–1194. <http://dx.doi.org/10.1016/j.jclinepi.2010.04.011>
- Collins, F. S., & Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine*, 372(9), 793–795. <http://dx.doi.org/10.1056/nejmp1500523>
- Dang, J. H., Rodriguez, E. M., Luque, J. S., Erwin, D. O., Meade, C. D., & Chen, M. S., Jr. (2014). Engaging diverse populations about biospecimen donation for cancer research. *Journal of Community Genetics*, 5(4), 313–327. <http://dx.doi.org/10.1007/s12687-014-0186-0>
- Dash, C., Wallington, S. F., Muthra, S., Dodson, E., Mandelblatt, J., & Adams-Campbell, L. L. (2014). Disparities in knowledge and willingness to donate research biospecimens: A mixed-methods study in an underserved urban community. *Journal of Community Genetics*, 5(4), 329–336. <http://dx.doi.org/10.1007/s12687-014-0187-z>
- Dodd, M., Janson, S., Facione, N., Faucett, J., Froelicher, E. S., Humphreys, J., et al. (2001). Advancing the science of symptom management. *Journal of Advance Nursing*, 33(5), 668–676.
- Dudley, J. T., Listgarten, J., Stegle, O., Brenner, S. E., & Parts, L. (2015). Personalized medicine: From genotypes, molecular phenotypes and the quantified self, towards improved medicine. *Pacific Symposium in Biocomputing*, 20, 342–346.
- Fox, S., & Duggan, M. (2013). Tracking for health. *Pew Research Center's Internet & American Life Project*. Retrieved from <http://www.pewinternet.org/2013/01/28/tracking-for-health/>
- Ghafoori, B., Barragan, B., Tohidian, N., & Palinkas, L. (2012). Racial and ethnic differences in symptom severity of PTSD, GAD, and depression in trauma-exposed, urban, treatment-seeking adults. *Journal of Traumatic Stress*, 25(1), 106–110. <http://dx.doi.org/10.1002/jts.21663>

- Glickman, S. W., Anstrom, K. J., Lin, L., Chandra, A., Laskowitz, D. T., Woods, C. W., et al. (2008). Challenges in enrollment of minority, pediatric, and geriatric patients in emergency and acute care clinical research. *Annals of Emergency Medicine*, 51(6), 775–780.e3. <http://dx.doi.org/10.1016/j.annemergmed.2007.11.002>
- Gwadz, M. V., Colon, P., Ritchie, A. S., Leonard, N. R., Cleland, C. M., Riedel, M., et al. (2010). Increasing and supporting the participation of persons of color living with HIV/AIDS in AIDS clinical trials. *Current HIV/AIDS Report*, 7(4), 194–200. <http://dx.doi.org/10.1007/s11904-010-0055-3>
- Hansen, M. M., Miron-Shatz, T., Lau, A. Y., & Paton, C. (2014). Big data in science and healthcare: A review of recent literature and perspectives. Contribution of the IMIA Social Media Working Group. *Yearbook of Medical Informatics*, 9(1), 21–26. <http://dx.doi.org/10.15265/iy-2014-0004>
- IBM. (2015). *IBM big data & analytics hub*. Retrieved from <http://www.ibmbigdatahub.com/infographic/four-vs-big-data>
- Kahn, J. P., Vayena, E., & Mastroianni, A. C. (2014). Opinion: Learning as we go: Lessons from the publication of Facebook's social-computing research. *Proceedings of the National Academy of Sciences U S A*, 111(38), 13677–13679. <http://dx.doi.org/10.1073/pnas.1416405111>
- Koffman, J., Gao, W., Goddard, C., Burman, R., Jackson, D., Shaw, P., et al. (2013). Progression, symptoms and psychosocial concerns among those severely affected by multiple sclerosis: A mixed-methods cross-sectional study of Black Caribbean and White British people. *PLoS One*, 8(10), e75431. <http://dx.doi.org/10.1371/journal.pone.0075431>
- Lee, Y. J., Boden-Albala, B., Larson, E., Wilcox, A., & Bakken, S. (2014). Online health information seeking behaviors of Hispanics in New York City: A community-based cross-sectional study. *Journal of Medical Internet Research*, 16(7), e176. <http://dx.doi.org/10.2196/jmir.3499>
- Lupton, D. (2013). *The commodification of patient opinion: The digital patient experience economy in the age of big data*. Sydney, Australia: Sydney Health & Society Group.
- Marr, B. (2015). *Big Data: The 5 Vs*. Retrieved February, 2015 from <http://www.slideshare.net/BernardMarr/140228-big-data-volume-velocity-variety-varacity-value>
- Martinez, K. A., Snyder, C. F., Malin, J. L., & Dy, S. M. (2014). Is race/ethnicity related to the presence or severity of pain in colorectal and lung cancer? *Journal of Pain and Symptom Management*, 48(6), 1050–1059. <http://dx.doi.org/10.1016/j.jpainsymman.2014.02.005>
- McSweeney, J. C., O'Sullivan, P., Cleves, M. A., Lefler, L. L., Cody, M., Moser, D. K., et al. (2010). Racial differences in women's prodromal and acute symptoms of myocardial infarction. *American Journal of Critical Care*, 19(1), 63–73. <http://dx.doi.org/10.4037/ajcc2010372>
- Merriman, J. D., Aouizerat, B. E., Cataldo, J. K., Dunn, L. B., Kober, K., Langford, D. J., et al. (2014). Associations between catecholaminergic, GABAergic, and serotonergic genes and self-reported attentional function in oncology patients and their family caregivers. *European Journal of Oncology Nursing*. <http://dx.doi.org/10.1016/j.ejon.2014.11.004>
- The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The belmont report: Ethical principles and guidelines for the protection of human subjects of research*. Bethesda, MD: The Commission.
- National Center for Advancing Translational Sciences. (2015). *Accural to clinical trials network*. Retrieved February 28, 2015, from <http://www.ncats.nih.gov/news-and-events/features/ctsa-act.html>
- National Institutes of Health. (2015). *NIH Big Data to Knowledge (BD2K)*. Retrieved February 28, 2015, from <http://bd2k.nih.gov/index.html#sthash.8SsjsH39.dpbs>
- National Institute of Nursing Research. (2015). *Spotlight on symptom management*. Retrieved February 28, 2015 from <https://www.ninr.nih.gov/researchandfunding/symptommanagement#VPM-DvnF8QE>
- Patient-Centered Outcomes Research Institute. (2015). *PCORnet: The national patient-centered clinical research network*. Retrieved February 28, 2015, from <http://www.pcornet.org/>

- Pickett, Y. R., Greenberg, R. L., Bazalais, K. N., & Bruce, M. L. (2014). Depression treatment disparities among older minority home healthcare patients. *American Journal of Geriatric Psychiatry*, 22(5), 519–522. <http://dx.doi.org/10.1016/j.jagp.2013.01.078>
- Reame, N. (2013). Research ethics. In A. Peirce & J. Smith (Eds.), *Ethical and legal issues for doctoral nursing students* (pp. 33–78). Lancaster, PA: Destech.
- Reddy, S. Y., Rasmussen, N. A., Fourie, N. H., Berger, R. S., Martino, A. C., Gill, J., et al. (2014). Sleep quality, BDNF genotype and gene expression in individuals with chronic abdominal pain. *BioMed Central Medical Genomics*, 7(1), 61. <http://dx.doi.org/10.1186/s12920-014-0061-1>
- Shaibi, G. Q., Coletta, D. K., Vital, V., & Mandarino, L. J. (2013). The design and conduct of a community-based registry and biorepository: A focus on cardiometabolic health in Latinos. *Clinical and Translational Science*, 6(6), 429–434. <http://dx.doi.org/10.1111/cts.12114>
- Sittig, D. F., Hazlehurst, B. L., Brown, J., Murphy, S., Rosenman, M., Tarczy-Hornoch, P., et al. (2012). A survey of informatics platforms that enable distributed comparative effectiveness research using multi-institutional heterogeneous clinical data. *Medical Care*, 50(Suppl), S49–S59. <http://dx.doi.org/10.1097/mlr.0b013e318259c02b>
- Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data*, 1(2), 85–99. <http://dx.doi.org/10.1089/big.2012.0002>
- Uddin, M., Koenen, K. C., Aiello, A. E., Wildman, D. E., de los Santos, R., & Galea, S. (2011). Epigenetic and inflammatory marker profiles associated with depression in a community-based epidemiologic sample. *Psychology Medicine*, 41(5), 997–1007. <http://dx.doi.org/10.1017/s0033291710001674>
- Van Den Eeden, S. K., Shan, J., Jacobsen, S. J., Aaronsen, D., Haque, R., Quinn, V. P., et al. (2012). Evaluating racial/ethnic disparities in lower urinary tract symptoms in men. *Journal of Urology*, 187(1), 185–189. <http://dx.doi.org/10.1016/j.juro.2011.09.043>
- Van Sickle, D., Magzamen, S., Truelove, S., & Morrison, T. (2013). Remote monitoring of inhaled bronchodilator use and weekly feedback about asthma management: An open-group, short-term pilot study of the impact on asthma control. *PLoS One*, 8(2), e55335. <http://dx.doi.org/10.1371/journal.pone.0055335>
- Vayena, E., Salathe, M., Madoff, L. C., & Brownstein, J. S. (2015). Ethical challenges of big data in public health. *PLoS Computational Biology*, 11(2), e1003904. <http://dx.doi.org/10.1371/journal.pcbi.1003904>
- Velez, M., Bales, M. E., Arcia, A., & Bakken, S. (2014). Electronic Tailored Infographics for Community Engagement, Education, and Empowerment (EnTICE3). *AMIA Joint Summits Translational Science Proceedings, eCollection 2014*, p. 206.
- Wikipedia. (2015). *Data science*. Retrieved February 28, 2015 from http://en.wikipedia.org/wiki/Data_science
- Wilcox, A., Fort, D., & Bakken, S. (2014). Creating a next-generation research informatics infrastructure: WICER lessons for data integration. *AMIA Joint Summits Translational Science Proceedings, eCollection 2014*, p. 167.
- Wilcox, A. B., Gallagher, K. D., Boden-Albala, B., & Bakken, S. R. (2012). Research data collection methods: From paper to tablet computers. *Medical Care*, 50(Suppl), S68–S73. <http://dx.doi.org/10.1097/mlr.0b013e318259c1e7>
- Yoon, S., Elhadad, N., & Bakken, S. (2013). A practical approach for content mining of Tweets. *American Journal of Preventive Medicine*, 45(1), 122–129. <http://dx.doi.org/10.1016/j.amepre.2013.02.025>
- Yoon, S., Wilcox, A., & Bakken, S. (2013). Comparisons among health behavior surveys: Implications for the design of informatics infrastructures that support comparative effectiveness research. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, 1(1), Article 9. <http://dx.doi.org/10.13063/2327-9214.1021>