SHORT
COMMUNICATION

Advancing Cancer Prevention and Behavior Theory in the Era of Big Data

Audie A. Atienza[1], Katrina J. Serrano[2], William T. Riley[3], Richard P. Moser[2], William M. Klein[2]

[1]ICF, Rockville, MD, [2]National Cancer Institute, Rockville, MD, [3]National Institutes of Health, Bethesda, MD, USA

The era of "Big Data" presents opportunities to substantively address cancer prevention and control issues by improving health behaviors and refining theoretical models designed to understand and intervene in those behaviors. Yet, the terms "model" and "Big Data" have been used rather loosely, and clarification of these terms is required to advance the science in this area. The objectives of this paper are to discuss conceptual definitions of the terms "model" and "Big Data", as well as examine the promises and challenges of Big Data to advance cancer prevention and control research using behavioral theories. Specific recommendations for harnessing Big Data for cancer prevention and control are offered.
(J Cancer Prev 2016;21:201-206)

Key Words: Cancer, Prevention, Data set, Health behavior

INTRODUCTION

Cancer remains a leading cause of death in the USA[1] and worldwide.[2] Improving health behaviors, such as smoking cessation, physical activity, eating a healthful diet, and adherence to evidenced-based cancer screening guidelines, remain key strategies in the prevention and control of cancer.[3] It has been argued that systematically examining the basis of human behaviors, guided by theory, can significantly enhance our understanding of cancer-related health behaviors and help design programs to improve these behaviors.[4] In support of this argument, prior research in the behavioral sciences has noted that health behavior interventions based on explicit theoretical models are more effective at changing the specific behaviors compared with interventions that are not theoretically based.[5]

Reviews of empirical research have revealed that only a fraction of published health behavior interventions have actually used theory to develop their respective interventions.[6] Of the limited interventions that do incorporate theory, most are based on a small number of general behavioral theories originally developed more than 30 years ago, are often informed by or loosely based on theory, and focus primarily on the individual level of analysis rather than potential influences at multiple levels (e.g., environmental and policy levels).[6] Furthermore, most studies that claim to be theory-based do not actually measure the theoretical constructs that are proposed as being responsible for behavior change, and a significant amount of variance remains unexplained.[7] These themes of limited use of theory in behavior change interventions and poor implementation of theory in intervention research when theory is actually used are reflected in the cancer literature more specifically.[8] Further advancement in theory development for behavioral change is needed to substantively move the field of cancer prevention and control forward.

Our understanding of human behaviors and ways to change cancer-relevant health behaviors can be substantively advanced

by utilizing and analyzing the massive amounts of data, referred to as *Big Data*, afforded by technological advances in the research enterprise. In this paper, we discuss promising opportunities and methodological approaches relevant to Big Data for cancer prevention and control issues in three areas: 1) data mining activities; 2) testing current theories with Big Data; and 3) integrating research models and methods from other fields into the behavioral sciences. This paper does not provide a comprehensive review of all studies relevant to Big Data and cancer prevention, but instead emphasizes future directions to advance cancer prevention and control by improving behavioral theory through the use of Big Data.

MATERIALS AND METHODS

The National Cancer Institute (NCI) organized a workshop - "Big Data and Theory Advancement" - held September 2013 at the National Institutes of Health (Bethesda, MD, USA). Experts in cancer prevention, computer engineering, statistics, behavioral science, and public health gathered to discuss how to leverage Big Data and dynamic systems models to advance health behavior theory in the context of cancer research. Workshop discussions and breakout groups were organized around the opportunities and challenges within five thematic topic areas: health behavior theory, systems modeling, social network data analysis, Big Data mash-ups and statistical modeling, and dynamic interventions. This paper reflects and expands on key themes and ideas discussed during this workshop.

RESULTS

1. Defining "model" and "Big Data"

We first make a distinction among three types of *models* and discuss definitions of *Big Data*, given that these terms are used and defined differently in various scientific fields. A *conceptual model* represents proposed relationships among constructs and is often graphically represented as a series of boxes indicating constructs, as well as arrows that demonstrate the relations among these constructs. Conceptual models need not be explanatory or predictive in nature, but when they are, they are often described as a "theory" in the health behavior field. In contrast, a *statistical model*, widely used by behavioral scientists, represents mathematical relationships among measures of constructs (typically discussed in terms of the direction and strength of association). While statistical models are often utilized to test hypotheses (e.g., multi-level regression modeling), many researchers have utilized statistical models to describe relationships among constructs without incorporating a theory or conceptual model. A *computational model*, used more in the engineering and computer science fields, also represents mathematical relationships of constructs. With computational models, however, researchers manipulate parameters of a complex, often dynamic system, using extensive computational resources (i.e., computer science tools) in an experimental manner (i.e., computer simulation) to make precise mathematical predictions.

The term *Big Data* has been used to represent various types of data (e.g., genomic data, social media data, real-time wearable sensor or cell phone data), often without clearly defining the term. For the purpose of this paper, Big Data is defined along three dimensions: volume, velocity, and variety.[9] Big Data differs from traditional data not only in the amount of data collected (i.e., volume) but also in how rapidly and efficiently large amounts of data can be collected, extracted, aggregated and/or integrated (i.e., velocity), often into more complex data sets. Big Data may be unstructured or structured text, number, image, audio and/or video (i.e., variety) and often must be "cleaned" or manipulated before it can be useful. Examples of Big Data relevant to cancer prevention and control include minute-by-minute accelerometry, global positioning system, and/or heart rate data on a large group of users; social media sites with millions of data points related to health behaviors; databases that merge large cohort studies with common data elements relevant to cancer; and large data sets derived from electronic health records and/or personalized health records with cancer-related information. With these definitions in mind, we discuss opportunities to leverage Big Data to advance cancer prevention and control research.

2. Data mining: Big Data and cancer prevention

Data mining can be very useful for generating and/or refining hypotheses by finding associations or patterns in large data sets that may not have otherwise been identified. Data mining is not one method but consists of a family of methods including decision trees, nonlinear regression and classification methods, and neural networks.[10] When appropriately used, data mining methods are interactive and iterative in nature. They involve selecting a relevant database, knowing the content of the data, performing data cleaning before any analyses, and choosing algorithms to examine relationships among variables in the data.

While data mining approaches are not new, so far only a few studies have employed these methods for the purpose of addressing topics relevant to cancer prevention and control. To date, most data mining studies of cancer-relevant behaviors from

mobile applications (apps) and/or social media platforms[11] have primarily been descriptive in nature. Fortunately, examples exist in the literature of how to conduct data mining on a very large sample with cross-sectional observation data to identify systematic correlates of cancer prevention outcomes, and rapidly validate the exploratory findings.[12] Moreover, an analytic framework exists for employing data mining methods with intervention studies (e.g., randomized clinical trial [RCT]).[13] Yet, employing data mining methods and corresponding validation analyses with longitudinal Big Data, in either repeated assessment observation or experimental studies, remains an unexplored frontier. Such explorations could help researchers identify new time-specific predictors of cancer-related health behaviors and contribute to the development of new behavioral theories or to the refinement of existing behavioral theories.

Machine learning, where computer algorithms can learn from and make predictions on data, also holds promise for behavior theory development because of its focus on prediction and its requirement for users to supply specific 'inputs' to be examined. Although relevant applications of this method can be found in the examination of genetic data to predict clinical outcomes,[14] there are few, if any, examples of this method being applied to predicting cancer-related health behaviors, much less behavior theory development using large data sets. Instead, cancer prevention-related studies employing machine learning[15] (e.g., natural language processing) have primarily been descriptive in nature, rather than predictive.

There are a number of limitations and challenges to mining Big Data, in general, that apply to cancer prevention and control research. There still exist many barriers to accessing Big Data, and even when accessible, there may be concerns about the quality of data, partly due to a lack of standard formats for data storage and linkage. A lack of behavioral ontologies also impedes progress by not providing standard definitions of constructs nor delineating relationships among constructs. Moreover, there is the concern that Big Data may have substantial 'noise' or errors, and thus do not have any veracity or true value. On a related note, there is concern that researchers may make inappropriate inferences or report spurious associations due to the nature of data-driven analyses. Replication of results to demonstrate robust findings[12] and knowledge synthesis to build a cumulative scientific database may help to address these concerns.

The proliferation of interactive internet sites, social media platforms (e.g., Facebook, Twitter, YouTube), smart phone apps (e.g., MyFitnessPal, QuitStart), and other mobile health wearable devices (e.g., Fitbit, Apple Watch, Garmin) have created potential data mining opportunities not previously conceived as possible. Collaborations with social media and health app companies to analyze de-identified datasets relevant to cancer prevention topics could help advance the field. In addition, the inclusion of cancer-relevant health behaviors (e.g., smoking status, cancer screening) in electronic health records as core objectives of Meaningful Use Stage 2, as discussed by the Institute of Medicine,[16] offers the possibility of accessing and analyzing large clinical datasets to understand and predict these key cancer-related behaviors. The ability to pool data from multiple data sets with common data elements and conduct integrative data analysis[17] with the larger combined data set offers further opportunities to explore cancer prevention and control issues.

3. Testing existing theories with Big Data

Big Data affords opportunities for directly testing and refining existing theories used in cancer prevention research, integrating them where appropriate, and discarding theories or parts of theories that are not empirically supported. In observational and quasi-experimental research, new technologies (e.g., mobile phones, sensors, social media) are being used to capture rich, temporally dense measurements (multiple observations/person/day) of health behavior and theoretical constructs in unprecedented detail to examine within- and between-person variability. These technologies expand the range of constructs that can be incorporated into new theories of health behavior by assessing the context of behavior in ways not previously possible. This also captures more precisely the timing of events, allowing for more detailed knowledge about their temporal ordering. For example, research using real-time mobile phone assessments has shown morning levels of self-efficacy, but not outcome expectancies, to predict leisure time physical activity later in the day among endometrial cancer survivors.[18] In addition to mobile technologies, social media platforms (e.g., Facebook or Twitter) are gaining increased attention among researchers interested in behavioral interventions.[19] Yet, much of this prior research has been limited to relatively small convenience samples. The use of very large sample sizes or very large time-intensive data sets to directly examine health behavior theories relevant to cancer prevention and control is on the near horizon.

As observational and quasi-experimental studies often have limitations in establishing causality, RCTs have come to be accepted as the gold standard research design for evaluating whether a behavioral intervention or treatment "works".[20] The great expense and long duration of RCTs create pressure to design behavioral interventions as "packages" that bundle together as

many theoretically active intervention components as possible in hopes that the eventually completed trial will yield a significant treatment effect. Recent advances in adaptive experimental design, such as the Multiphase Optimization Strategy (MOST), the sequential multiple assignment randomized trial (SMART), and the micro-randomization study, allow optimization of behavioral interventions and refinement of behavioral theory using a RCT design.[21] While behavioral researchers have begun employing adaptive intervention designs, the use of Big Data in cancer prevention and control interventions has received scant attention, much less the testing of theory. Further investigations of how to incorporate these novel optimized RCT designs into theory testing with very large samples are warranted.

Distinct from traditional and optimized RCTs, advances in and proliferation of mobile phone and sensor technologies provide opportunities for Just-in-Time, Adaptive Interventions (JITAIs). JITAIs are contextualized interventions provided at the place and time that they are needed and adapt to changes in individual behavior and needs.[22] Cancer prevention research is beginning to utilize JITAIs. For example, one pilot study found that JITAI reduced sedentary behavior among obese adults.[23] In another study, a Mobile TEEN smart phones app automatically detects physical activity and sedentary bouts, as well as prompts users to assess real-time theory-based predictors of these behaviors via time-intensive monitoring.[24] Further development of JITAIs promises rich sources of time-intensive Big Data to help researchers better understand and modify behavior tailored specifically to each individual.

Taken together, several opportunities hold promise for testing existing behavior theories relevant to cancer prevention and control. 1) Technology platforms (e.g., Fitbit, Apple Watch, Run-Keeper, etc.) that collect time-intensive observation behavior data could help advance theory by incorporating selected measures based on theoretical constructs. 2) Researchers can leverage mobile technology and/or social media to developing large-scale adaptive interventions to test whether the manipulation and optimization of various proposed theoretical factors (e.g., extrinsic motivation, self-efficacy) actually changes cancer-related health behaviors (e.g., smoking cessation). 3) Passive assessment of behaviors and environments via mobile and/or environmental sensors (e.g., accelerometers, passive smoking sensors, GPS) offer new opportunities for theory-based JITAIs tailored specifically for the individual.

4. Integrating models and methods to advance theory

Borrowing and adapting research methods and statistical/com-putational models from fields outside of behavioral science that address dynamic data may provide new avenues to advance cancer prevention and control, and radically transform how we test and refine health behavior theory. One profound change in data collection is the proliferation of temporally dense data from various technologies. These new sources of data for theoretical testing, however, require methods and analytic techniques that are designed to handle temporally dense, often noisy data.

Fortunately, many of these approaches already exist, predominantly from computer science and engineering where researchers address noisy, temporally dense data, leading to the development of robust and sophisticated methods for analyzing and modeling such data.[25] The field of health behavior theory has also begun to borrow from computer science and engineering a range of computational dynamic modeling approaches, generally termed systems science models. Social network analysis, agent-based modeling, and dynamical systems modeling are the three major forms of computational modeling that have increasingly been used to study behavioral phenomena.[26] These computational approaches not only offer a greater mathematical specificity of the relations among theoretical constructs than statistical modeling but also provide substantial flexibility to model complex and dynamic interrelations among theoretical constructs over time.

We briefly describe the three forms of computational modeling. Social network analysis examines social influences via nodes (individual actors) and ties (the connection between nodes). Social network analysis has been used to characterize the influences of individuals on one another for a variety of health behaviors.[27] Agent-based models use computational models to simulate the dynamic actions of agents (individuals or collective groups such as corporations). In the behavioral and social sciences, agent-based models have been used primarily to understand the effects of population-based health policies (e.g., changes in cigarette taxes, increased access to immunization),[28] but could be used to address a wide variety of health outcomes and their antecedents. Dynamic system models represent a set of computational modeling approaches to model complex systems over time. Dynamic systems models stemming from control systems engineering have recently been applied to health behavior.[29] In modeling of feedback loops and the use of fluid analogies, these models can explain how even seemingly simple systems can behave in complex and nonlinear ways. The application and adaptation of computer science and engineering computational dynamic modeling approaches to the development of novel dynamic health behavior models have been discussed in

relation to behavioral research in general[21] and cancer-related health behaviors, such as tobacco interventions.[30]

However, the testing of these new dynamic models using large-scale Big Data to address cancer-related health behaviors has not received much attention. As such, the evidence of how well these dynamic models can capture the experiences of cancer-relevant populations, and the complex relationships between theoretical constructs and particular health behaviors is only preliminary. It also remains unclear how these new dynamic models correspond to or improve the traditional "static" models related to cancer-related health behavior change or health behavior links to cancer outcomes.

DISCUSSION

In this age of Big Data, traditional study designs (e.g., cross-sectional surveys, RCTs) and traditional research methods (e.g., simple regressions, pre- to post-intervention analyses) seem insufficient to capture the richness of the data that can now be collected for cancer prevention and control using Big Data sources. To substantively improve our understanding of cancer-related health behaviors and make modification to these key behaviors, further advancement of the theories that explain these behaviors are needed.

The following recommendations are put forth to advance cancer-related behavioral theories with Big Data:

1) Encourage data mining in all aspects of cancer prevention research, from data exploration aimed at hypothesis generation to intervention research aimed at refining hypotheses (e.g., post RCT exploration of treatment effects using CART). Establish training opportunities in data mining and data visualization approaches for behavioral scientists interested in cancer prevention and control research.

2) Develop, curate, and incorporate passive and/or brief computer-adaptive measures of cancer-related health behaviors and their proposed theoretical predictors into various studies and platforms that can collect a large amount of data (e.g., electronic health records, social media, mobile health apps, large cohort studies). Establish common data elements, common measures, and behavioral ontologies for cancer prevention researchers to use. Prioritize research that incorporates these common measures, and explicitly test proposed mechanisms of behavior change.

3) Encourage collaborations among cancer prevention researchers, data scientists, psychometric experts, computer engineers, clinical informatics researchers, bioinformatics experts, behavioral methodologists, and behavioral theorists to advance cancer prevention research and related theories. Funding opportunities, developer challenges/prizes, hackathons, symposia, and workshops could facilitate the formation of these collaborations.

4) Establish public-private partnerships that involve cancer prevention and control researchers working with health technology companies, social media companies, health app entrepreneurs, EHR vendors, and/or non-governmental organizations to collect information on cancer prevention relevant topics. The partnerships could emphasize the analysis of existing data, incorporating relevant measures into established or developing infrastructure, create application program interfaces to readily share data for analysis, and/or establish new methods and approaches for testing and refining behavioral theories.

5) Create proof-of-principle studies for implementing adaptive and optimized behavioral interventions in large cancer-relevant samples (e.g., Facebook cancer groups, health maintenance organization networks, online cancer communities). Explore the utilization of large cancer-related volunteer panels to accelerate the pace of behavioral intervention development and implementation via novel technology.

6) Compare head-to-head JITAI versus traditional/usual care behavioral interventions to evaluate the effectiveness of improving specific cancer-related behaviors. Measures of proposed theoretical mechanisms in both types of interventions should be included and explicitly tested.

7) Analyze dynamic models of behavior change relevant to cancer, and test whether dynamic models better explain behavior change (i.e., account for more variance) than traditional health behavior models.

To reduce the burden of cancer from a population science perspective, changing human behavior is essential. Armed with Big Data, health information technology, and rigorous research methodology, emerging innovations in research offer much promise to the scientific community in ever important endeavors to better understand and modify cancer-related health behaviors.

ACKNOWLEDGMENTS

of Southern California), Bob Evans (Google), Ross Hammond, PhD (The Brookings Institute), Eric Hekler, PhD (Arizona State University), Stephen Intille, PhD (Northeastern University), Holly Jimison, PhD (Northeastern University), Misha Pavel, PhD (Northeastern University), Daniel Rivera, PhD (Arizona State University), Alex Rothman, PhD (University of Minnesota), Bonnie Spring, PhD (Northwestern University), Donna Spruijt-Metz, PhD (University of Southern California), and Jasmin Tiro, PhD (University of Texas, Southwestern) along with a number of Federal agency participants.

CONFLICTS OF INTEREST

No potential conflicts of interest were disclosed.

The views and opinions expressed in this paper do not necessarily reflect those of the NCI, National Institutes of Health, or Department of Health and Human Services.

REFERENCES

1. National Center for Health Statistics. Health, United States, 2013, with special feature on prescription drugs. Hyattsville, MD, Government Printing Office, 2015.
2. World Health Organization. The top 10 causes of death. July 2013. http:// www.who.int/mediacentre/factsheets/fs310/en/. Accessed July 2014.
3. Hiatt RA, Rimer BK. A new strategy for cancer control research. Cancer Epidemiol Biomarkers Prev 1999;8:957-64.
4. Glanz K, Rimer BK. Theory at a glance: a guide for health promotion practice. 2nd ed. Bethesda, MD, National Cancer Institute, 2005.
5. Noar SM, Benac CN, Harris MS. Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. Psychol Bull 2007;133:673-93.
6. Glanz K, Bishop DB. The role of behavioral science theory in development and implementation of public health interventions. Annu Rev Public Health 2010;31:399-418.
7. Ogden J. Some problems with social cognition models: a pragmatic and conceptual analysis. Health Psychol 2003;22:424-8.
8. Kobrin S, Ferrer R, Meissner H, Tiro J, Hall K, Shmueli-Blumberg D, et al. Use of health behavior theory in funded grant proposals: cancer screening interventions as a case study. Ann Behav Med 2015;49:809-18.
9. Dumbill E. Getting up to speed with big data. In: Big Data Now. 2nd ed. Sebastopol, CA, O'Reilly Media, Inc., pp 3-17, 2012.
10. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference and prediction. New York, Springer-Verlag New York, pp 83-85, 2009.
11. Thackeray R, Burton SH, Giraud-Carrier C, Rollins S, Draper CR. Using twitter for breast cancer prevention: an analysis of breast cancer awareness month. BMC Cancer 2013;13:508.
12. Serrano KJ, Yu M, Coa KI, Collins LM, Atienza AA. Mining health app data to find more and less successful weight loss subgroups.

J Med Internet Res 2016;18:e154.
13. King AC, Ahn DF, Atienza AA, Kraemer HC. Exploring refinements in targeted behavioral medicine intervention to advance public health. Ann Behav Med 2008;35:251-60.
14. Kureshi N, Abidi SS, Blouin C. A predictive model for personalized therapeutic interventions in non-small cell lung cancer. IEEE J Biomed Health Inform 2016;20:424-31.
15. Chou WY, Prestin A, Kunath S. Obesity in social media: a mixed methods analysis. Transl Behav Med 2014;4:314-23.
16. Institute of Medicine. Capturing social and behavioral domains and measures in electronic health records: phase 2. Washington, DC, The National Academies Press, 2014.
17. Curran PJ, Hussong AM. Integrative data analysis: the simultaneous analysis of multiple data sets. Psychol Methods 2009;14:81-100.
18. Basen-Engquist K, Carmack CL, Li Y, Brown J, Jhingran A, Hughes DC, et al. Social-cognitive theory predictors of exercise behavior in endometrial cancer survivors. Health Psychol 2013;32:1137-48.
19. Kent EE, Prestin A, Gaysynsky A, Galica K, Rinker R, Graff K, et al. "Obesity is the new major cause of cancer": connections between obesity and cancer on facebook and twitter. J Cancer Educ 2016;31:453-9.
20. Carey MP, Senn TE, Coury-Doniger P, Urban MA, Vanable PA, Carey KB. Optimizing the scientific yield from a randomized controlled trial (RCT): evaluating two behavioral interventions and assessment reactivity with a single trial. Contemp Clin Trials 2013;36:135-46.
21. Riley WT, Rivera DE. Methodologies for optimizing behavioral interventions: introduction to special section. Transl Behav Med 2014;4:234-7.
22. Patrick K, Intille SS, Zabinski MF. An ecological framework for cancer communication: implications for research. J Med Internet Res 2005;7:e23.
23. Thomas JG, Bond DS. Behavioral response to a just-in-time adaptive intervention (JITAI) to reduce sedentary behavior in obese adults: Implications for JITAI optimization. Health Psychol 2015;34 Suppl:1261-7.
24. Dunton GF, Dzubur E, Kawabata K, Yanez B, Bo B, Intille S. Development of a smartphone application to measure physical activity using sensor-assisted self-report. Front Public Health 2014;2:12.
25. Kumar S, Nilsen WJ, Abernethy A, Atienza A, Patrick K, Pavel M, et al. Mobile health technology evaluation: the mHealth evidence workshop. Am J Prev Med 2013;45:228-36.
26. Luke DA, Stamatakis KA. Systems science methods in public health: dynamics, networks, and agents. Annu Rev Public Health 2012;33:357-76.
27. Valente TW, Gallaher P, Mouttapa M. Using social networks to understand and prevent substance use: a transdisciplinary perspective. Subst Use Misuse 2004;39:1685-712.
28. Elliott E, Kiel LD. Agent-based modeling in the social and behavioral sciences. Nonlinear Dynamics Psychol Life Sci 2004;8:121-30.
29. Navarro-Barrientos JE, Rivera DE, Collins LM. A dynamical model for describing behavioural interventions for weight loss and body composition change. Math Comput Model Dyn Syst 2011;17:183-203.
30. Timms KP, Rivera DE, Collins LM, Piper ME. Continuous-time system identification of a smoking cessation intervention. Int J Control 2014;87:1423-37.