TBM

CASE STUDY

CYberinfrastructure for COmparative effectiveness REsearch (CYCORE): improving data from cancer clinical trials

Kevin Patrick, MD, MS,[1,2,3] Laura Wolszon, PhD,[3] Karen M Basen-Engquist, PhD, MPH,[4] Wendy Demark-Wahnefried, PhD, RD,[5] Alex V Prokhorov, MD, PhD,[4] Stephanie Barrera, MS, RD, LD,[4] Chaitan Baru, PhD,[6] Emilia Farcas, PhD,[7] Ingolf Krueger, PhD,[7] Doug Palmer, PhD,[3] Fred Raab,[1,2] Phil Rios,[3] Celal Ziftci,[7] Susan Peterson, PhD[4]

[1]Department of Family and Preventive Medicine, UC San Diego, Calit2, 9500 Gilman Drive, MC 0811, La Jolla, CA 92093-0811, USA

[2]Center for Wireless and Population Health Systems, UC San Diego, Calit2, 9500 Gilman Drive, MC 0811, La Jolla, CA 92093-0811, USA

[3]UC San Diego, Calit2, 9500 Gilman Drive, MC 0811, La Jolla, CA 92093-0811, USA

[4]Department of Behavioral Science, The University of Texas, MD Anderson Cancer Center, P.O. Box 301439, Houston, TX 77030, USA

[5]UAB Comprehensive Cancer Center, NP 2514, 1530 3rd Ave. S., Birmingham, AL 35294-3360, USA

[6]UC, San Diego Supercomputing Center, 9500 Gilman Drive, MC 0505, La Jolla, CA 92093-0505, USA

[7]Department of Computer Science and Engineering, UC San Diego, 9500 Gilman Drive, MC 0404, La Jolla, CA 92093-0404, USA

Correspondence to: K Patrick
kpatrick@ucsd.edu

Cite this as: *TBM* 2011;1:83–88
doi: 10.1007/s13142-010-0005-z

ABSTRACT

Improved approaches and methodologies are needed to conduct comparative effectiveness research (CER) in oncology. While cancer therapies continue to emerge at a rapid pace, the review, synthesis, and dissemination of evidence-based interventions across clinical trials lag in comparison. Rigorous and systematic testing of competing therapies has been clouded by age-old problems: poor patient adherence, inability to objectively measure the environmental influences on health, lack of knowledge about patients' lifestyle behaviors that may affect cancer's progression and recurrence, and limited ability to compile and interpret the wide range of variables that must be considered in the cancer treatment. This lack of data integration limits the potential for patients and clinicians to engage in fully informed decision-making regarding cancer prevention, treatment, and survivorship care, and the translation of research results into mainstream medical care. Particularly important, as noted in a 2009 report on CER to the President and Congress, the limited focus on health behavior-change interventions was a major hindrance in this research landscape (DHHS 2009). This paper describes an initiative to improve CER for cancer by addressing several of these limitations. The Cyberinfrastructure for Comparative Effectiveness Research (CYCORE) project, informed by the National Science Foundation's 2007 report "Cyberinfrastructure Vision for 21st Century Discovery" has, as its central aim, the creation of a prototype for a user-friendly, open-source cyberinfrastructure (CI) that supports acquisition, storage, visualization, analysis, and sharing of data important for cancer-related CER. Although still under development, the process of gathering requirements for CYCORE has revealed new ways in which CI design can significantly improve the collection and analysis of a wide variety of data types, and has resulted in new and important partnerships among cancer researchers engaged in advancing health-related CI.

KEYWORDS

Comparative effectiveness research (CER), Information technology (IT), Neoplasms, Health behavior, Clinical trials, Environmental monitoring, Home-monitoring

Implications

Practice: Clinicians will find decision-making easier and more effective for the patients because new variables can be taken into account, including those arising from individual differences. A physician can then prescribe treatment regimens with more confidence of positive outcomes.

Policy: The information gleaned from better-supported CER studies can be used to inform national health policy, and because of increased effectiveness, can lower the cost of health care significantly for patients, providers and insurers.

Research: CYCORE facilitates the collection and analysis of home-based physiological, behavioral, social and environmental data from patients undergoing cancer treatment. Largely either unmeasured or self-reported, these data are essential to the quality and applicability of cancer-related clinical trials.

Adherence to medical regimens has been appreciated historically as a key factor that influences outcomes of clinical trials in cancer. Recent data also suggest that lifestyle factors such as tobacco and alcohol use, physical activity, dietary consumption, and energy balance may serve as powerful prognostic indicators of the outcome of clinical trials as well as long-term success of cancer treatment [2–7]. Additionally, a growing body of literature is demonstrating the impact of environmental factors on disease outcomes. Even though low-cost and increasingly ubiquitous technologies support objective measurement in each of these domains–treatment adherence, health behavior, and continuing environmental exposures–these findings often remain unaccounted for in comparative effectiveness research (CER) because of the challenges inherent in collecting, processing, and acting upon this

ever-increasing amount of data. This problem is exacerbated by the need in CER to merge these data with information from other sources such as medical records, physiological monitors, and patient self-reports.

CER, with its emphasis on overall patient outcomes and quality of life rather than short-term biological or clinical endpoints, may stand to benefit greatly from more powerful approaches to handle these data in health decision-making. However, few models exist of systems designed specifically to integrate and interpret the variety of data important to cancer CER. This paper provides an overview of one such system in development, CYberinfrastructure for COmparative Effectiveness REsearch (CYCORE): improving data from cancer clinical trials and outlines some of the challenges encountered in its initial phases.

CYCORE is being developed by collaborators with backgrounds in behavioral science, clinical research, and information technology (IT). It involves designing a prototype of an IT system that supports the acquisition, storage, visualization, analysis, and sharing of data acquired within and across clinical trials. A particular emphasis is being placed upon ensuring that CYCORE has the ability to incorporate data from diverse sources, in a variety of formats and over multiple studies, with a special focus on acquisition of behavioral, lifestyle, and environmental data. Since future studies might gather data that are now unanticipated, the infrastructure for CYCORE must be scalable, easily modified, and adaptable to changing requirements, data structures, maintainability, and governance.

Figure 1 shows a simplified view of the CYCORE system and how stakeholders will interact with it. Stakeholder categories include research participants (e.g., cancer patients or survivors, family members, and community members), researchers and research teams, oncologists and other health-care providers, resource providers (including tool developers and medical data providers), operators, and policy makers. The aim is to serve researchers and health-care providers through a system that supports (1) data collection, assimilation, and quality assurance from existing data sources (e.g., electronic medical records) and emerging data sources (sensor measurements and self-reports), (2) data distribution and event notification for use by researchers, clinicians, policy makers, and operators of the cyberinfrastructure (CI), (3) data analysis and visualization capabilities, and (4) governed service and resource integration to address security and privacy concerns.

Development of CYCORE is centered on five areas of effort:

1) Obtaining system requirements from cancer researchers, clinicians, and patients.

To ensure that CYCORE is easily usable and can be integrated into current CER, requirements and preferences are being gathered from clinicians, researchers, resource providers, and research participants. This includes compiling information about specific health-related states that require monitoring, and the methods for doing so. Requirements have been defined in the context of specific use-cases that are based on actual research protocols. In addition to information about sources of data, requirements are being gathered about the forms of analyses that are needed and for what purposes (e.g., intervention or generation of reports), the appearance and functionality of patient and provider interfaces, how the resources are managed, the activities the system should support, and data-

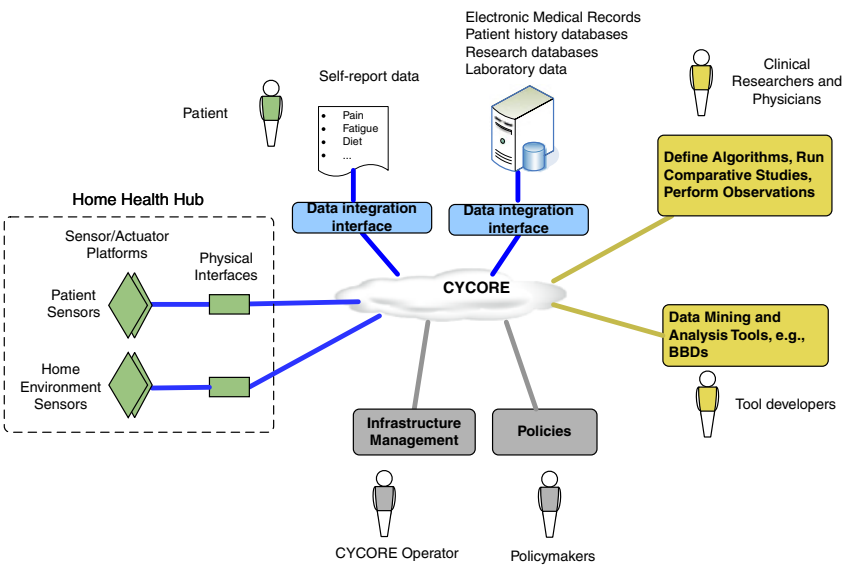

Fig 1 | CYCORE scenario. Data from patient self-reports and body-worn sensors (interacting with the Home Health Hub, see below), complemented by fixed sensors in the environment, can be collected and analyzed (using, for example, Brain-Based Devices) to perform Comparative Effectiveness studies

security needs (i.e., standards, policies, Health Information Portability and Accountability Act requirements, etc.).

Stakeholders were identified from a single institution (MDACC) using both purposeful and snowball sampling methods. To date, requirements have been collected from 61 stakeholders representing most of the key stakeholder categories identified *a priori* for CYCORE, including research investigators and research staff (n=20), oncology physicians, and other health-care providers (n=16), administrators and policy makers (n=10), and informatics/IT experts (n=15). The methodology for requirements elicitation followed an iterative process identifying stakeholder roles, their goals, their activities and the context in which they operate, limitations in the current protocols, features expected from the CYCORE system (e.g., data acquisition from patient sensors), and quality-of-service requirements (e.g., usability, portability, and reliability). The methodology employed domain modeling (see Figs. 2 and 3), i.e., creation of models depicting issues of particular importance for any given type of cancer and its treatment, and their relationships. Establishing a common understanding and language is key for eliciting requirements and involves acquiring domain knowledge, observing the environment, and establishing trust relationships with all parties.

User scenarios were detailed by the stakeholders, showing how each would interact with the system to achieve a particular goal (e.g., how a researcher wants to design and monitor a study, how a patient would like to follow his or her progress in regard to specific monitoring outcomes, etc.). In addition to requirements from researchers and clinicians, several meetings were held with providers of electronic solutions for existing medical and research data, with the goal of defining integration points with CYCORE.

The requirements-elicitation process is iterative. Thus, an initial list of requirements was identified, which advances the architecture of the system and the implementation of a prototype, and then a new set of requirements and refinements is generated via multiple feedback loops. New requirements were then reviewed and prioritized so as to include them in the next iteration of the prototype, and enables validation of their implementation. This iterative process for requirements engineering helps reduce the risks of designing an inadequate system or of over-engineering.

Several themes have emerged from these exchanges including the need for CYCORE to (1) convey frequently updated symptom assessment and intervention-adherence feedback to clinicians and researchers, (2) automate the collection of questionnaire-type dietary and quality-of-life data, (3) provide algorithms for analysis and validation of data from multiple sensors (e.g., the integration of information from global positioning systems and accelerometers as a measure of physical functioning), (4) integrate institutionally established ontologies with more globally accepted structures (e.g., the cancer Biomedical Informatics Grid [caBIG]), and (5) continually reevaluate what is doable versus what cannot be accomplished given institutional and technology-related constraints.

2) Creating a system that enables multi-format data aggregation, integration, processing, mining, storage and retrieval.

CYCORE is leveraging systems-level data integration, processing, mining, and storage technologies that are already being developed by collaborators from UCSD's Calit2 and San Diego Supercomputer Center, with support from the National Science Foundation, the NIH, and others. These are being further enhanced to deal with the specific data types and application requirements obtained from CYCORE's initial panel of users. CYCORE is being developed as open-source software and will have a service-oriented "Rich

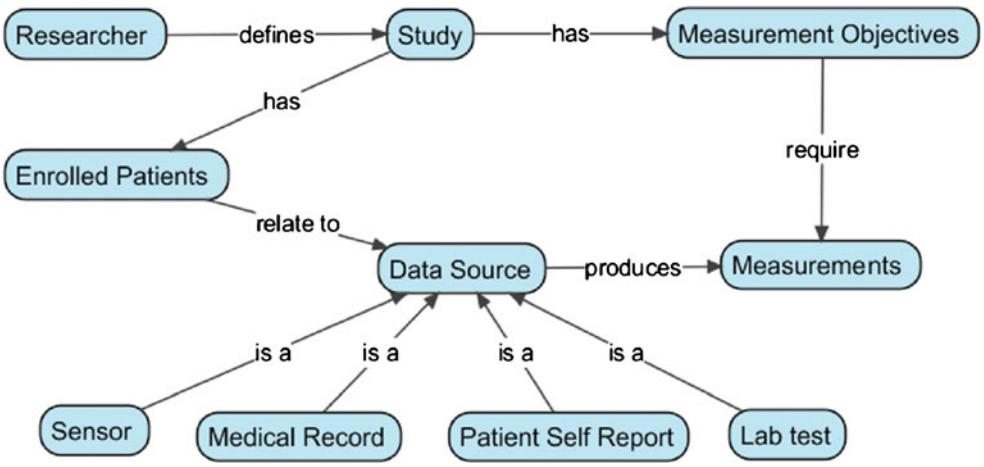

Fig 2 | Example of a domain model depicting a simplified view of a study

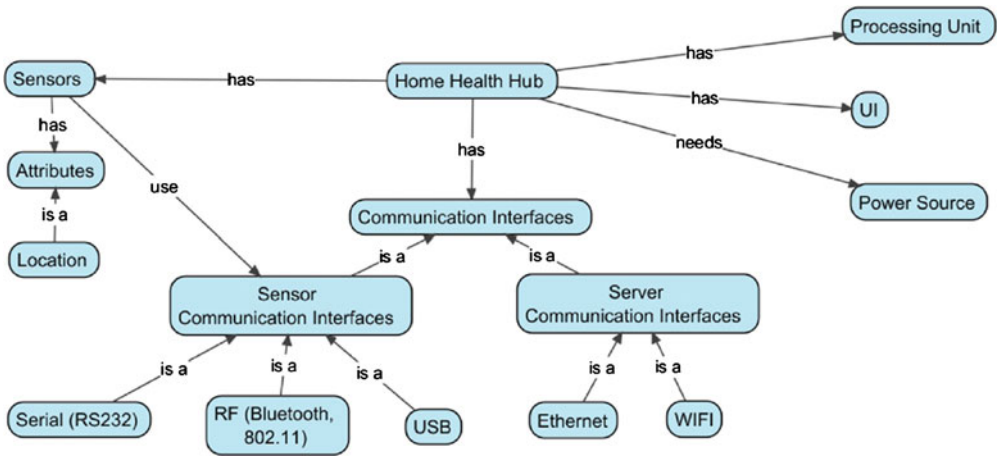

Fig 3 | Example of a domain model depicting a simplified view of the Home Health Hub (HHH) sensor platform

Service" architecture (described below) designed to integrate complex information obtained from a variety of data types and sources, including electronic medical records, sensing and imaging devices, and other large-scale IT support systems in oncology such as the caBIG.

CYCORE is a cyber-physical system, that is, a system combining physical entities such as sensors and mobile devices with processes such as data acquisition, using an underlying computational and data infrastructure (an integrated set of hardware and software including user interfaces, middleware, servers, and networks). This CI will provide security, dependability, maintainability, scalability, flexibility, and other important properties. For example, electronic medical systems have a set of additional challenges related to data privacy, which also must be addressed.

To facilitate the integration of system capabilities and features into a scalable infrastructure, CYCORE is being built as a service-oriented architecture (SOA), a paradigm of software development that provides "loose coupling" (i.e., weak dependencies) between its various system services (Fig. 4). Services are the mechanism by which specific needs (e.g., a researcher's need to visualize data in a web browser) and capabilities (e.g., obtaining data from two different sensors) are

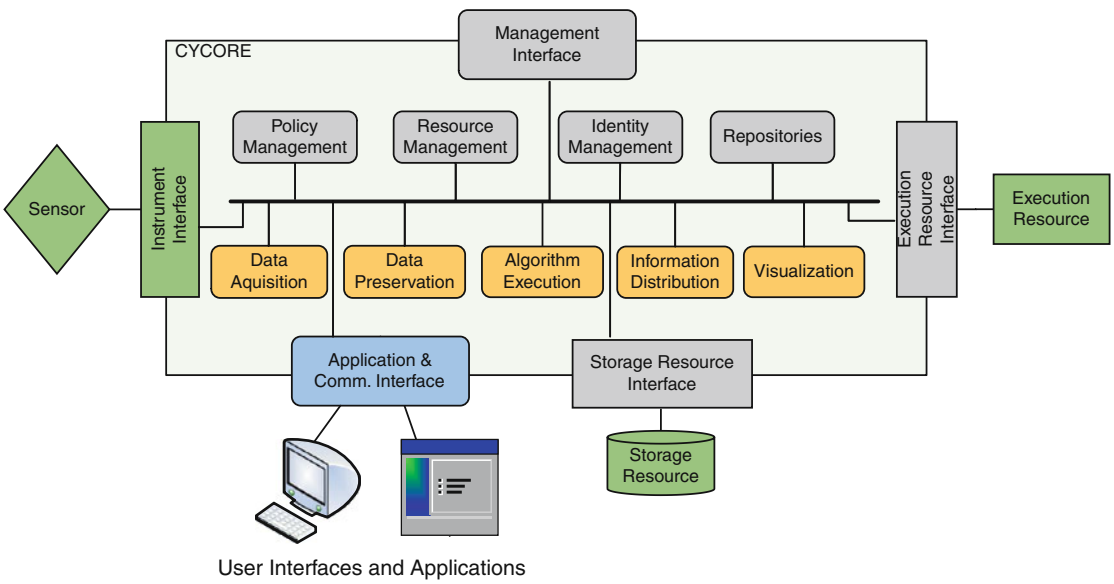

Fig 4 | Cyberinfrastructure (CI) interfaces and services. The CI will provide interfaces to sensors, data and storage resources, and user applications. The CI provides core capabilities such as data acquisition, preservation, distribution, visualization, and algorithm execution. The CI also manages crosscutting concerns such as policy and identity management within the infrastructure

brought together within a single software system [13]. Thus, the SOA approach promotes independent development and reuse of software components that, in turn, reduce development cost. SOAs provide the means to offer, discover, and interact with CYCORE's functions (e.g., acquiring sensor data or storing and exporting data to software packages normally used for CER outcomes assessment).

All functional capabilities and resources are being represented as services in the CYCORE system, with precisely defined service-access protocols. Moreover, CYCORE uses a Rich Service Architecture [8], which is a type of SOA for organizing complex systems and managing distributed capabilities that may be under the control of different ownership domains. For example, with multiple users, each institution may have its own policy for the usage of its resources and data. Rich Services allow for infrastructure services, such as policy management (e.g., authorization, privacy, and auditing) and identity management (authentication and management of user identities), to be plugged into the architecture without modifying core system functionality. This feature ensures scalability, so CYCORE can grow as new needs are identified, and new users engage with the system without changes to the underlying CI.

As described earlier and depicted in the figures, CYCORE consists of physical resources such as sensors, storage resources such as disks and network drives, and larger servers necessary for high-performance computing interfacing with increasingly common "cloud computing", data servers allocated on-demand on shared infrastructures that are increasingly offered by entities such as Microsoft and Amazon. CYCORE also includes a research–participant interface for the Home Health Hub (HHH)—the physical device that integrates all patient data acquired in the home—as well as tools and applications for researchers. Data mining and analysis tools such as Brain-Based Devices (see item 4 below) are application modules that can be plugged into the infrastructure as desired, through an application interface.

CYCORE poses data-related challenges that go beyond traditional data-management systems including: (1) end-to-end data preservation (long-term storage of all data in the system) and access, (2) management of associations between data collected from various sources, but related to the same patient (e.g., medical records, patient sensors, environmental sensors, etc.), (3) execution of data-analysis algorithms that themselves produce new data to be ingested in the system and used by collaborators, (4) complex event-based processing to detect certain events in the studies (such as detecting dehydration risk in patients), and (5) time-stamping of data as it is routed through the system. Further, CYCORE addresses semantic heterogeneities across data sources where, for example, different terms may actually refer to the same concept [9].

Because the SOA disentangles infrastructure and the applications that use it, policies no longer clutter the specifications of each individual application. Rather, they are off-loaded to and implemented in the infrastructure itself, which then applies the policies where and when needed in each application. This leads to significant flexibility and scalability of the system, as well as increased confidence in the adherence to policies: one only needs to go to a single location in the architecture and the source code to see whether a given policy is followed.

3) Building and/or integrating data-acquisition hardware and software.

The primary interface between the research subject and the CI will be provided via CYCORE's HHH, to which the biometric and environmental sensors transmit data. Originally developed at Calit2, the HHH aggregates sensor data and relays these to the CYCORE data management system via a web-service interface. The HHH consists of a small computer with physical interfaces supporting radio-, serial- and USB-enabled sensors. Thus far, monitoring capabilities of the HHH via wireless devices include blood pressure, weight, physical activity, photo/video, environmental tobacco smoke, and medication adherence via "smart pill bottles."

The HHH enables two-way communication between patients and researchers through audio and video feeds and can support interfaces such as touch-screens on which questionnaires for self-report data can be displayed. CYCORE collects data in sensor streams that are compatible with those from other data sources important to cancer CER, such as clinical records and lab data. Within the last 18 months, there has been a rapid increase in the range and extent of inexpensive sensors of all types that can be deployed via wireless connection to the HHH. Also, several systems that have some of the features of the HHH have come on the commercial market. Evaluation of how these systems might support CYCORE is presently underway.

4) Data analysis with a "Brain-Based Device" analytic system.

Collecting data on patient treatments and outcomes must be matched with a strong ability to compile and interpret the wide range of variables that must be considered in the cancer treatment. One of the most innovative elements of CYCORE will be to explore the use of a Brain-Based Device (BBD) to assist with the processing and interpretation of the complex data sets collected in CER studies. BBDs were developed at the Neurosciences Institute of La Jolla under the direction of Nobel-laureate Gerald Edelman [10–12] and were then advanced from concept to desktop implementation by the Intellisis Corporation in conjunction with Calit2 researchers. BBDs are beginning to change

how very large amounts of disparate data can be processed through use of a neurobiology-inspired approach to information analysis, categorization, and correlation because they have the unique ability to accept missing and corrupted data without disturbing the primary "learning" trends in the data mining, and have been utilized successfully in robotics and smart systems.

The advantage that BBDs have over traditional statistical data-mining techniques is their ability to comprehend, model, and make inferences from very large numbers of data inputs and outputs. CYCORE will explore the use of a BBD to emulate the responses of patients to cancer treatments. Variables will include those that relate to patient history, cancer treatment, and behavioral and environmental factors in the course of treatment. Through "computer training" of the BBD, we will attempt to model patients' responses and thus predict the success of future courses of treatment. While this 2-year project will not be able to incorporate this technology to its fullest extent, evaluation of its feasibility and usability in this project will set the stage for future studies.

5) Conducting pilot studies of the use of CYCORE.

After developing the first (stripped-down) iteration of CYCORE, we will assess its functionality and ease-of-use use by testing data from a sample of cancer survivors and their caregivers or family members. Through ongoing feasibility testing, we will evaluate the processes of data collection, transfer, and integration. Simple web interfaces will be customized for various stakeholder groups (patients, cyber-operators, policy makers, and researchers). Information gleaned from this evaluation will enable improved iterations with increasing functionality and permit us to plan the course of subsequent stages of development.

Exploiting existing resources for evaluating user-centered designs (e.g., www.usability.gov), we will establish milestones for determining feasibility, such as patients' responses to prompts for self-reported data, activation and communication of home- or personal monitoring equipment, patient adherence to procedures for self-monitoring and/or self-reported data capture, successful uploading and storage of data, and ability to search and access data for analysis.

Outcomes will be assessed objectively according to customary protocols. Field staff will also conduct debriefing interviews with participants to assess their experiences using the HHH suite of devices and to troubleshoot any problems or barriers to their use that might arise. Data collected in the debriefing will immediately be summarized and reported to the investigators to ensure rapid problem resolution.

CONCLUSION

The vision for the CYCORE project is to develop a prototype of a comprehensive, state-of-the-art IT-based system to enable large-scale and robust CER across the cancer continuum, i.e., from cancer prevention to cancer treatment and ultimately, to cancer control and survivorship care. The need for such an infrastructure is great given the complexities inherent in the prevention and treatment of cancer. To be most meaningful, questions in CER should be informed by not only medical and clinical data but data from other domains critical to health such as behaviors and the everyday experiences of individuals in their home environment. Subsequent reports from this project will provide further insight into how to accomplish this and should guide others as they endeavor to translate research into practice and improve patient and population-level outcomes.

Acknowledgements: This study is supported by the National Institutes of Health (National Cancer Institute), grant # RC2CA148263-01.

1. DHHS (U.S. Department of Health and Human Services) (2009). *Federal Coordinating Council for Comparative Effectiveness Research. Report to the President and Congress*. 30 Jun 2009.
2. Gritz, E. R., Dresler, C., & Sarna, L. (2005). Smoking, the missing drug interaction in clinical trials: ignoring the obvious. *Cancer Epidemiology, Biomarkers & Prevention, 14*, 2287–2293.
3. Bairati, I., Meyer, F., Jobin, E., et al. (2006). Antioxidant vitamins supplementation and mortality: A randomized trial in head and neck cancer patients. *International Journal of Cancer, 119*, 2221–2224.
4. Chlebowski, R. T., Aiello, E., & McTiernan, A. (2002). Weight loss in breast cancer patient management. *Journal of Clinical Oncology, 20*, 1128–1143.
5. Holmes, M. D., Chen, W. Y., Feskanich, D., Kroenke, C. H., & Colditz, G. A. (2005). Physical activity and survival after breast cancer diagnosis. *JAMA, 293*, 2479–2486.
6. Meyerhardt, J. A., Heseltine, D., Niedzwiecki, D., et al. (2006). Impact of physical activity on cancer recurrence and survival in patients with stage III colon cancer: findings from CALGB 89803. *Journal of Clinical Oncology, 24*, 3535–3541.
7. Velicer, C. M., & Ulrich, C. M. (2008). Vitamin and mineral supplement use among US adults after cancer diagnosis: a systematic review. *Journal of Clinical Oncology, 26*, 665–673.
8. Arrott, M., Demchak, B., Ermagan, V., Farcas, C., Farcas, E., Krüger, I. H., et al. (2007). *Rich services: the integration piece of the SOA puzzle, in proceedings of the IEEE International Conference on Web Services (ICWS)* (pp. 176–183). Salt Lake City, Utah, USA: IEEE.
9. Baru, C., & Lin, K. (2009). *Mediating among GeoSciML resources*. In *International Journal of Digital Earth* (pp. 18–28). London: Taylor & Francis.
10. Krichmar, J. L., & Edelman, G. M. (2005). Brain-based devices for the study of nervous systems and the development of intelligent machines. *Artificial Life, 11*, 63–77.
11. Edelman, D. B., Baars, B. J., & Seth, A. K. (2005). Identifying hallmarks of consciousness in non-mammalian species. *Consciousness and Cognition, 14*, 169–187.
12. Seth, A. K. (2005). Causal connectivity of evolved neural networks during behavior. *Network, 16*, 35–54.
13. OASIS Reference Model for Service Oriented Architecture 1.0 http://docs.oasis-open.org/soa-rm/v1.0/soa-rm.pdf, 2006