

Big Data for Population-Based Cancer Research: The Integrated Cancer Information and Surveillance System

Anne-Marie Meyer, Andrew F. Olshan, Laura Green, Adrian Meyer, Stephanie B. Wheeler, Ethan Basch, William R. Carpenter

The Integrated Cancer Information and Surveillance System (ICISS) facilitates population-based cancer research by developing extensive information technology systems that can link and manage large data sets. Taking an interdisciplinary “team science” approach, ICISS has developed data, systems, and methods that allow researchers to better leverage the power of big data to improve population health.

Advances in information technology and data computing have revolutionized population-based research by providing access to large quantities of secondary data; these linked databases are sometimes referred to as “big data” [1]. Using these data for public health research or clinical research requires that researchers broaden their horizons beyond the traditional surveillance model, as working with big data is very different than analyzing narrowly focused, treatment-oriented clinical trial data. Big data must be carefully and effectively leveraged if it is to accurately reflect the heterogeneous populations it represents. This effort requires an agile research environment that quickly adopts advances in computing technology to continually integrate data while applying novel methods to untangle their complexity.

North Carolina’s Integrated Cancer Information and Surveillance System (ICISS) is a novel research data system enabled by unprecedented support from the North Carolina General Assembly through the University Cancer Research Fund [2]. ICISS employs an interdisciplinary “team science” approach that requires the close collaboration of researchers from many fields, including clinical and population sciences, as well as unparalleled partnerships with computer science and health informatics professionals [3]. Integrating and operationalizing these data requires significant investments in a secure, high-level, data-computing and research infrastructure that is nimble enough to address longitudinal data needs and can adapt rapidly to evolving research methods.

The mission of ICISS is to improve cancer outcomes in North Carolina by assembling, linking, and harmonizing big data to facilitate high-impact, cancer-focused research spanning the cancer continuum [4]. This goal is accomplished through integrated activities relating to data sets, systems, and methods (see Figure 1).

First, ICISS develops and maintains an innovative, comprehensive, and prospectively linked library of large population-based data sets that include measures across the cancer care continuum, from screening to postdiagnosis outcomes. Second, ICISS includes a secure virtual computing platform and a software development team, which together deliver innovative research tools and meet technical needs. Integrated into daily work flows, these tools enable navigation of clinical coding catalogs, cohort discovery, project tracking, and knowledge retention. Third, ICISS facilitates cutting-edge cancer outcomes research by cultivating an interdisciplinary team environment and by applying novel data management and analytic methods for studying large sets of nonexperimental data. Many aspects of the ICISS system, including its unique coding search, tracking tools, and methods expertise, are publicly accessible to North Carolina cancer researchers at <http://iciss.unc.edu>.

ICISS Governance and Data Security

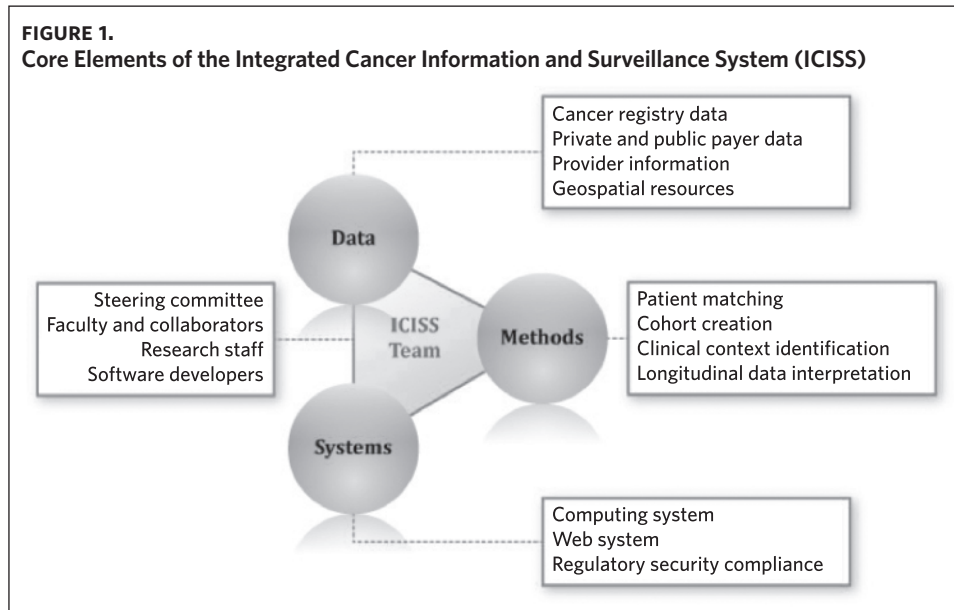
A tightly managed set of governance policies and procedures helps to ensure that use of ICISS resources aligns with the research missions of ICISS and of the University Cancer Research Fund, and that use of these resources is in compliance with numerous security safeguards and regulatory requirements. The governance process is overseen by a multidisciplinary steering committee of researchers who represent several schools within the University of North Carolina (UNC) and various data partners, including the North Carolina Central Cancer Registry (NCCCR). This oversight spans all steps of the research process, from an initial letter proposing a research study using the ICISS data, to security training and oversight of projects, to prepublication review of research products. The UNC Institutional Review Board (IRB) oversees protocols for maintaining and linking ICISS data as a resource, as well as for each research study that uses ICISS data. In addition, a global security management

Electronically published July 2, 2014.

Address correspondence to Dr. Anne-Marie Meyer, 101 E Weaver St, CB #7293, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7293 (meyera@email.unc.edu).

N C Med J. 2014;75(4):265-269. ©2014 by the North Carolina Institute of Medicine and The Duke Endowment. All rights reserved. 0029-2559/2014/75408

FIGURE 1.
Core Elements of the Integrated Cancer Information and Surveillance System (ICISS)



plan preserves the confidentiality, integrity, and availability of personally identifiable information and protected health information, including limited data sets. This security plan ensures regulatory compliance with the Health Insurance Portability and Accountability Act of 1996 (HIPAA), the Federal Information Security Management Act of 2002, and North Carolina's state law regarding data security breaches [5].

ICISS Data

The integrated ICISS research data have given North Carolina a unique ability to perform sophisticated analyses in multiple, diverse research areas including comparative effectiveness, health disparities, quality improvement, patient-centered outcomes research, public health services and systems research, implementation science, epidemiologic and biostatistical methods, health information technology, health economics, and organizational studies. The data incorporate numerous measures on multiple levels and characterize a wide range of influential factors, including personal characteristics and behaviors, health care organizational factors, and broader environmental influences (see Figure 2). These measures span the cancer continuum from risk and genetic predisposition through diagnosis, treatment, intermediate outcomes, and end-of-life care.

Person-level data. Person-level data are drawn from numerous sources, including the NCCCR, Medicare, Medicaid, and private health insurance plans. When linked, these combined data can richly characterize all North Carolinians with cancer, as well as those without cancer, for the years 2003 through 2010. NCCCR data provide detailed information about the cancer diagnosis such as tumor size, aggressiveness, and extent of disease; basic demographic information; the date of diagnosis; vital status; and the date of death. Administrative and insurance claims data charac-

terize the health characteristics and health care utilization of cancer and noncancer patients, which is important for several reasons, including understanding of cancer prevention and early detection.

Cases are linked to multipayer claims data sets through deterministic and probabilistic linkage methods [6]. Together these data sets cover approximately 5.5 million unique individuals—about 55% of the state's overall population and about 70% of the cancer patients in North Carolina. Claims files provide information on diagnoses, procedures, and dates of service, and they allow researchers to observe patients before and during diagnosis, through treatment, and during continued follow-up. Treatment morbidity (eg, toxicities) and comorbidities can be observed for as long as the patient maintains insurance coverage and receives billable health services. Mortality and date of death are captured within the NCCCR data and can often be verified using the claims enrollment files. The data are updated regularly as more information becomes available from relevant data partners.

Provider-level data. Extensive research has demonstrated the influence of health care provider characteristics on cancer diagnosis, treatment, and outcomes. ICISS data include information from the North Carolina Health Professions Data System (managed by UNC's Cecil G. Sheps Center for Health Services Research) [7], the American Medical Association, and the American Hospital Association, and this information can be linked to claims data. ICISS also has geocoded data from the North Carolina Health Professions Data System (2000–2010), which allows researchers to examine the impact of distance to care and other access issues [8–10]. [Editor's note: For an example of how ICISS data were used to study the impact of access to care, see the original article by Wheeler and colleagues on pages 239–246]. In addition, these data can be used for geospatial analyses of environmen-

tal and other factors. All provider data are managed centrally within ICISS and are thoroughly deidentified after linkage, so identifying data are not released to research teams.

Area-level data. Area-level data sets can be linked at multiple geographic levels to characterize environmental, socioeconomic, and political contexts [11, 12]. These data sets include Area Health Resources Files from the Health Resources and Services Administration [13], the Robert Wood Johnson Foundation's County Health Rankings [14], the *National Profile of Local Health Departments* compiled by the National Association of County and City Health Officials [15], the Behavioral Risk Factor Surveillance System of the Centers for Disease Control and Prevention, and US Census tract data from the American Community Survey. Together these data include thousands of variables that contextualize North Carolina counties and local health departments.

Reference data. The usability of observational data is contingent on the successful interpretation and use of coded data. For example, linking and using ICISS data require knowledge of US Census data; name(s) data; codes for medical diagnoses, procedures, and drugs; and crosswalks between codes, including city, ZIP, census, and county codes. ICISS has developed a unique Web-based reference search system that allows for high-validity, efficient code identification, definition, and crosswalk.

The ICISS clinical coding tool systematically normalizes and links various nomenclatures through multiple cross-

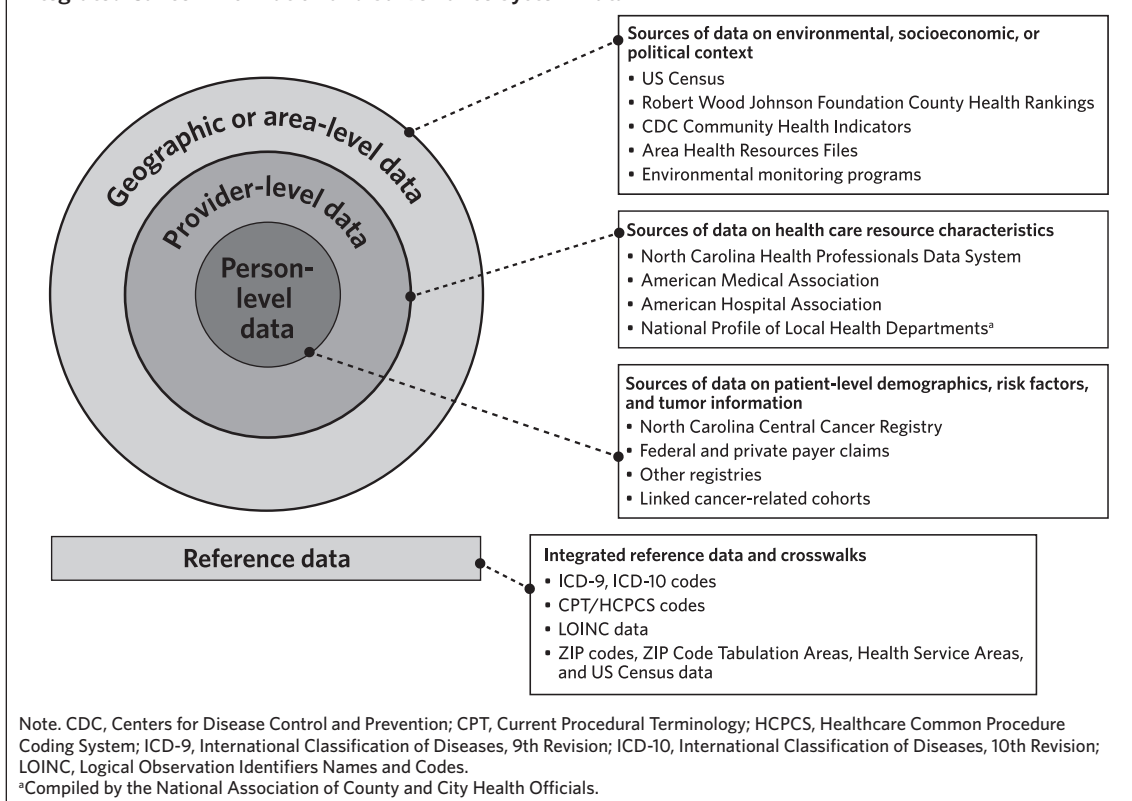
walks; these nomenclatures include the 9th and 10th revisions of the International Classification of Diseases (ICD-9 and ICD-10); the International Classification of Diseases for Oncology (ICD-O); Current Procedural Terminology (CPT); Healthcare Common Procedure Coding System (HCPCS); Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT); National Drug Codes (NDC); Anatomical Therapeutic Chemical (ATC) Classification System; and Logical Observation Identifiers Names and Codes (LOINC). The clinical coding tool helps researchers navigate catalogs by enabling advanced searching mechanisms through synonym matching. Users can group codes into meaningful concepts related to diagnoses, treatments, and outcomes specific to each study, and this information is retained upon completion of a study and made available through the Web-based system to inform future research.

ICISS Systems

Systems infrastructure—the second core element of ICISS—is aligned to support the entire research project life cycle while ensuring the security and privacy of ICISS data.

Aggregated data environment. An online platform for querying deidentified, aggregated data is designed to help investigators perform preliminary data review through tailored views of ICISS data. Variables such as cancer site, age, sex, and ethnicity can be overlaid with census data, health indicators, and socioeconomic variables to assess basic

FIGURE 2.
Integrated Cancer Information and Surveillance System Data



measures and study feasibility. The easy-to-navigate Web interface allows users to save reports, maps, or tables and to share them with other users.

Secure data analysis platform. ICISS operates a secure data analysis platform in collaboration with UNC Information Technology Services. This infrastructure enhances accessibility through remote virtual desktops, gets new users up and running quickly, delivers extremely fast data transfer, and centralizes licenses and installations of tools (eg, SAS, R, ArcGIS). Users connect to ICISS through 2-factor authentication. Secure data access is managed centrally at ICISS, and levels of access are individualized for each user based on the user's role, his or her Data Use Agreement, and IRB governance requirements.

Research tracking system. The central research tracking system promotes consistent study governance and allows project managers to oversee research operations and activities. It provides electronic capture and review of proposals, letters of intent, IRB applications, and grant proposals. It can also track the progress and deliverables of ongoing projects, including abstracts and publications, and it can assign tasks across a matrix team. The system is customized to triage and streamline requests for data access.

ICISS Research and Methods

An interdisciplinary "team science" approach is required for resources such as ICISS to operate efficiently and effectively [14, 16-18]. The ICISS team includes computer scientists, clinicians, biostatisticians, epidemiologists, health services researchers, demographers, and geographers trained at the MD, PhD, and master's degree levels. The team is fully integrated, and individuals work together to share diverse ideas and to develop and optimize novel interdisciplinary solutions for managing, leveraging, and lifting complex data sets into analytic files for research studies.

Research scope. The ICISS data and the ICISS team enable a diverse portfolio of research. ICISS is currently being used for studies of comparative effectiveness (eg, the article by Goyal and colleagues on pages 231-238), treatment disparities [10, 19], access to care [8], and investments within the public health system (eg, a 2011 study by Mays and Smith [20], which linked increases in public health spending to declines in preventable deaths). The availability of longitudinal data allows researchers to examine the effect of health policies such as Medicare Part D or the Patient Protection and Affordable Care Act of 2010, as well as temporal trends such as the 2008 economic recession [21]. Disease incidence, risk of late effects of treatment, and rare diseases can also be studied by applying novel geospatial statistical methods (eg, a 2011 study by Kuo and colleagues [22], which reported geographic disparities in late-stage breast cancer diagnosis).

Population-based and advanced methods. ICISS is uniquely positioned to support the development of new, advanced analytic methods that will extend the reach of its research

potential. For example, ICISS can be used to study the different types of bias that exist in observational studies [23]. Multipayer data allow researchers to examine the selection and confounding biases found in single-payer data (eg, Medicare) or clinical trials, as well as patient heterogeneity and differential patient outcomes [24-27]. ICISS data also enable exploration of instrumental variables for control of unmeasured confounding [28-31], causal modeling [32, 33], systems modeling (ie, agent-based models), or application of social network analysis (ie, care coordination [34]). Finally, ICISS is facilitating work in advanced data mining and data visualization by collaborating with computer scientists, informaticians, and biostatisticians.

Data integration. ICISS continues to develop novel methods of enhancing data, including 3-way and 4-way integration of epidemiologic cohort studies and other registries. This allows a more comprehensive characterization of the measures that are important to the study of cancer care and outcomes, and it substantially expands the research beyond what would be possible with any of the data sets alone. For example, the integrated data allow examination of critical individual-level measures of behavioral risk factors, laboratory results, patient-reported outcomes, and genetic markers—none of which are currently available in registry or claims data [35]. This is opening the door to more personalized medicine and patient-centered research, as well as the development of new methods for creating, managing, and using big data.

Conclusion

Effectively using big data for population research requires more than just access to terabytes or petabytes of data. Both technical and human resources are required to operationalize integrated research systems and environments for this type of data resource. Developing such resources requires transdisciplinary collaborations of well-trained professionals who are able to develop novel hypotheses, bridge technical and disciplinary gaps, and communicate effectively to achieve cohesive solutions. Agile and secure computing systems are needed to support the data and to meet the specific requirements of data partners and researchers. Although grant funding can offset some of the costs of personnel, a preliminary investment is essential for data acquisition and development of technical infrastructure and systems. With current data systems in place, a governance structure has been implemented to expand access to ICISS to investigators across the state. Developed through the support of the North Carolina General Assembly and the University Cancer Research Fund, ICISS represents a successful integrated research platform that leverages large, linked, multipayer data sets to improve population health. **NCMJ**

Anne-Marie Meyer, PhD research assistant professor, Department of Epidemiology, UNC Gillings School of Global Public Health, University of North Carolina at Chapel Hill; facility director, Integrated Cancer Information and Surveillance System (ICISS), Lineberger Comprehensive

Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

Andrew F. Olshan, PhD distinguished professor, Department of Epidemiology, UNC Gillings School of Global Public Health, University of North Carolina at Chapel Hill; department chair, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

Laura Green, MBA project manager, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

Adrian Meyer, MS director of systems development, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

Stephanie B. Wheeler, PhD, MPH assistant professor, Department of Health Policy and Management, UNC Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

Ethan Basch, MD, MSc director, Cancer Outcomes Research Program, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

William R. Carpenter, PhD, MHA faculty director, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill; associate professor, Department of Health Policy and Management, UNC Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

Acknowledgment

Potential conflicts of interest. All authors have no relevant conflicts of interest.

References

- Ward JS, Barker A. Undefined by data: a survey of big data definitions. arXiv:1309.5821v1 [csDB] <http://arxiv.org/abs/1309.5821>. Published September 20, 2013. Accessed April 1, 2014.
- O'Malley MS, Blouin R, Pisano ED, Rimer BK, Roper WL, Earp HS 3rd. Research for North Carolina: the University Cancer Research Fund. *N C Med J*. 2008;69(4):299-302.
- Stokols D, Hall KL, Taylor BK, Moser RP. The science of team science—overview of the field and introduction to the supplement. *Am J Prev Med*. 2008;35(2 suppl):S77-S89.
- Carpenter WR, Meyer AM, Abernethy AP, Stürmer T, Kosorok MR. A framework for understanding cancer comparative effectiveness research data needs. *J Clin Epidemiol*. 2012;65(11):1150-1158.
- The North Carolina Identity Theft Protection Act of 2005, amended in 2006. NCGS §75-65.
- Agency for Healthcare Research and Quality (AHRQ). Developing and evaluating methods for record linkage and reducing bias in patient registries. Effective Health Care Program. Topic abstract (research summary). AHRQ Web site. <http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?productid=1362&pageaction=displayproduct>. December 11, 2012. Accessed April 27, 2014.
- North Carolina Health Professions Data System. Cecil G. Sheps Center for Health Services Research Web site. <http://www.shepscenter.unc.edu/hp/>. Accessed April 1, 2014.
- Holmes JA, Carpenter WR, Wu Y, et al. Impact of distance to a urologist on early diagnosis of prostate cancer among black and white patients. *J Urol*. 2012;187(3):883-888.
- Stitzenberg KB, Chang Y, Louie R, Groves JS, Durham D, Fraher EF. Improving our understanding of the surgical oncology workforce. *Ann Surg*. 2014;259(3):556-562.
- Wheeler SB, Wu Y, Meyer AM, et al. Use and timeliness of radiation therapy after breast-conserving surgery in low-income women with early-stage breast cancer. *Cancer Invest*. 2012;30(4):258-267.
- Athens JK, Catlin BB, Remington PL, Gangnon RE. Using empirical Bayes methods to rank counties on population health measures. *Prev Chronic Dis*. 2013;10:E139. doi:0.5888/PCD10.130028.
- Peppard PE, Kindig DA, Dranger E, Jovaag A, Remington PL. Ranking community health status to stimulate discussion of local public health issues: the Wisconsin County Health Rankings. *Am J Public Health*. 2008;98(2):209-212.
- Health Resources and Services Administration (HRSA). Area Health Resources Files (AHRF): National, state and county health resources information database. HRSA Web site. <http://arh.hrsa.gov/>. Accessed April 1, 2014.
- County Health Rankings and Roadmaps. County Health Rankings Web site. <http://www.countyhealthrankings.org/>. Accessed April 1, 2014.
- National Association of County and City Health Officials (NACCHO). National Profile of Local Health Departments. NACCHO Web site. <http://www.naccho.org/topics/infrastructure/profile/resources/>. Accessed April 1, 2014.
- Leischow SJ, Best A, Trochim WM, et al. Systems thinking to improve the public's health. *Am J Prev Med*. 2008;35(2 suppl):S196-S203.
- Mabry PL, Olster DH, Morgan GD, Abrams DB. Interdisciplinary and systems science to improve population health: a view from the NIH Office of Behavioral and Social Sciences Research. *Am J Prev Med*. 2008;35(2 suppl):S211-S224.
- Stokols D, Misra S, Moser RP, Hall KL, Taylor BK. The ecology of team science: understanding contextual influences on transdisciplinary collaboration. *Am J Prev Med*. 2008;35(2 suppl):S96-S115.
- Carpenter WR, Tyree S, Wu Y, et al. A surveillance system for monitoring, public reporting, and improving minority access to cancer clinical trials. *Clin Trials*. 2012;9(4):426-435.
- Mays GP, Smith SA. Evidence links increases in public health spending to declines in preventable deaths. *Health Aff (Millwood)*. 2011;30(8):1585-1593.
- Wheeler SB, Kohler RE, Goyal RK, et al. Is medical home enrollment associated with receipt of guideline-concordant follow-up care among low-income breast cancer survivors? *Med Care*. 2013;51(6):494-502.
- Kuo TM, Mobley LR, Anselin L. Geographic disparities in late-stage breast cancer diagnosis in California. *Health Place*. 2011;17(1):327-334.
- Meyer AM, Wheeler SB, Weinberger M, Chen RC, Carpenter WR. An overview of methods for comparative effectiveness research. *Semin Radiat Oncol*. 2014;24(1):5-13.
- Committee on Comparative Effectiveness Research Prioritization; Board on Health Care Services; Institute of Medicine. Initial National Priorities for Comparative Effectiveness Research. Washington, DC: National Academies Press; 2009.
- Methodology Committee of the Patient-Centered Outcomes Research Institute (PCORI). Methodological standards and patient-centeredness in comparative effectiveness research: the PCORI perspective. *JAMA*. 2012;307(15):1636-1640.
- Sox HC, Goodman SN. The methods of comparative effectiveness research. *Annu Rev Public Health*. 2012;33:425-445.
- Luce BR, Kramer JM, Goodman SN, et al. Rethinking randomized clinical trials for comparative effectiveness research: the need for transformational change. *Ann Intern Med*. 2009;151(3):206-209.
- Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol Drug Saf*. 2010;19(6):537-554.
- Brookhart MA, Wyss R, Layton JB, Stürmer T. Propensity score methods for confounding control in nonexperimental research. *Circ Cardiovasc Qual Outcomes*. 2013;6(5):604-611.
- Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution—a simulation study. *Am J Epidemiol*. 2010;172(7):843-854.
- Newhouse JP, McClellan M. Econometrics in outcomes research: the use of instrumental variables. *Annu Rev Public Health*. 1998;19:17-34.
- Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10(1):37-48.
- Hernán M, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15(5):615-625.
- Pollack CE, Weissman GE, Lemke KW, Hussey PS, Weiner JP. Patient sharing among physicians and costs of care: a network analytic approach to care coordination using claims data. *J Gen Intern Med*. 2013;28(3):459-465.
- Meyer AM, Carpenter WR, Abernethy AP, Stürmer T, Kosorok MR. Data for cancer comparative effectiveness research: past, present, and future potential. *Cancer*. 2012;118(21):5186-5197.