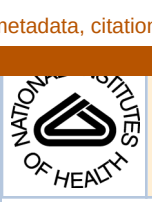

Published in final edited form as:

Cancer. 2012 November 1; 118(21): 5186–5197. doi:10.1002/cncr.27552.

Data for Cancer Comparative Effectiveness Research: Past, Present, and Future Potential

Anne-Marie Meyer, PhD[1,2], **William R Carpenter, PhD**[1,2,3], **Amy P. Abernethy, MD**[4,5], **Til Stürmer, MD PhD**[2,6], and **Michael R. Kosorok, PhD.**[1,7]

[1] UNC-Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC, USA.

[2] Cecil G. Sheps Center for Health Services Research, UNC, Chapel Hill, NC.

[3] Department of Health Policy and Management, Gillings School of Global Public Health, Chapel Hill, NC.

[4] Department of Medicine, Division of Medical Oncology, Duke University, Durham, NC.

[5] Duke Cancer Institute, Duke University, Durham, NC.

[6] Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC.

[7] Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC.

Abstract

Background—Comparative effectiveness research (CER) can efficiently and rapidly generate new scientific evidence and address knowledge gaps, reduce clinical uncertainty, and guide health care choices. Much of the potential in CER is driven by the application of novel methods to analyze existing data. Despite its potential, several challenges must be identified and overcome so that CER may be improved, accelerated, and expeditiously implemented into the broad spectrum of cancer care and clinical practice.

Methods—To identify and characterize the challenges to cancer CER, we reviewed the literature and conducted semi-structured interviews with 41 cancer CER researchers at the Agency for Healthcare Research and Quality (AHRQ)'s Developing Evidence to Inform Decisions about Effectiveness (DEcIDE) Cancer CER Consortium.

Results—A number of datasets for cancer CER were identified and differentiated into an ontology of eight categories, and characterized in terms of strengths, weaknesses, and utility. Several themes emerged during development of this ontology and discussions with CER researchers. Dominant among them was accelerating cancer CER and promoting the acceptance of findings, which will necessitate transcending disciplinary silos to incorporate diverse perspectives

Corresponding Author: William R Carpenter, PhD Department of Health Policy and Management University of North Carolina, Gillings School of Global Public Health 1102A McGavran Greenberg Hall; CB 7411 Chapel Hill, NC 27599 Phone 919-966-6328; Fax 919-966-6961 wrc4@email.unc.edu.

*The authors comprise the Data Committee of the Agency for Healthcare Research and Quality's Cancer DEcIDE Comparative Effectiveness Research Consortium: http://www.effectivehealthcare.ahrq.gov/index.cfm/who-is-involved-in-the-effective-health-care-program1/about-the-decide-network/

Disclaimer:
The views expressed in this article are those of the authors, and no official endorsement by the Agency for Healthcare Research and Quality or the U.S. Department of Health and Human Services is intended or should be inferred.

and expertise. Multidisciplinary collaboration is required including those with expertise in non-experimental data, outcomes research, clinical trials, epidemiology, generalist and specialty medicine, survivorship, informatics, data, and methods, among others.

Conclusions—Recommendations highlight the systematic, collaborative identification of critical measures; application of more rigorous study design and sampling methods; policy-level resolution of issues in data ownership, governance, access, and cost; and development and application of consistent standards for data security, privacy, and confidentiality.

INTRODUCTION

Rapid advances in cancer care continues through an accelerated pace of scientific discovery and technology development. Timely integration of developments into clinical practice is increasingly challenging, and it is imperative for more immediate, generalizable, and evidence-based information. Randomized controlled trials (RCTs) remain the gold standard for developing such information; however, this research design is not always feasible, practical, or sufficiently timely. Additionally, RCT designs limit generalizability of findings to heterogeneous patient populations or specific subgroups seen in clinical practice. [1-7]

Cancer comparative effectiveness research (CER) holds great promise for meeting many shortcomings of RCTs. Though CER takes many forms, for this discussion, we focus on the Institute of Medicine's (IOM) definition of CER:

> Comparative effectiveness research is the generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat, and monitor a clinical condition or to improve the delivery of care. The purpose of CER is to assist consumers, clinicians, purchasers, and policy makers to make informed decisions that will improve health care at both the individual and population levels.[6, 8]

The foundation of CER is understanding effectiveness in the context of large, heterogeneous populations. Propitiously, large population-based data are becoming increasingly available through advances in information technology and research methods in the form of secondary data collected for non-research purposes. By increasing our understanding of these data, CER stands to benefit immeasurably by these ever-growing repositories.

For cancer CER, these data originate from many different sources including electronic health records, registries, administrative data, observational studies, clinical trials, and others. Not all existing or secondary data are adequate, and each data source comes with its own unique challenges. Because secondary data originate from many different sources, they may be missing critical variables or have significant and systematic differences how variables are measured. . These differences impede the ability to confidently characterize important care processes and outcomes across data. An additional challenge is the lack of randomization, which makes controlling for relevant confounders critical. As a result, cancer care stakeholders are frequently uncomfortable acting on CER findings generated from these data sources.

A better understanding of data is necessary to improve data collection and methods development, to overcome the challenges facing cancer CER. To further this understanding, and help guide federal data and research partners, we reviewed the literature and met with over 40 cancer outcomes researchers and clinicians. Our goals were to: 1) develop a conceptual model for examining observational data in cancer CER; 2) characterize the strengths and limitations of current cancer CER data resources; 3) identify barriers in the conduct of cancer CER; and 4) formulate recommendations and guiding principles. While

our focus was on secondary, observational data (i.e., non-randomized, retrospective), the findings we present are also applicable to any prospective data collection.

METHODS

Data collection

Literature was reviewed regarding current cancer care, cancer research data, and cancer comparative effectiveness research. This information helped inform the development of a conceptual model of data needs for cancer CER,[9] and frame discussions with a convenience sample of cancer outcomes researchers associated with the Can-DEcIDE Consortium.[a] Participants were from multiple disciplines and included clinicians, clinical trials experts, epidemiologists, pharmacoepidemiologists, health services researchers, biostatisticians, clinical data managers, state public health workers, and informaticians. The majority of participants relied on federal or academic funding; individuals who relied on funding from industry or non-government third party payers were not targeted in the initial sampling frame. Applying snowball sampling, participants were asked to identify other researchers that may provide additional insight, and together comprised the study sample of 41 discussants.

Discussions were conducted individually and tailored according to each researcher's area of expertise. Guided by findings in the literature, discussions centered on the following: 1) identification of specific datasets for cancer CER; 2) utility of measures 3) data access or logistical challenges; 4) population/target and sampling; 5) data linking capabilities; 6) longitudinal follow-up in datasets; 7) temporality of data/measures; 8) data completeness; 9) data standardization, formatting, and documentation; and 10) data processing and required expertise.

Study Team Review and Development of Recommendations

The primary study team comprised an epidemiologist, pharmacoepidemiologist, biostatistician, health services researcher, and three cancer-focused physician researchers, all of whom conduct federally-funded patient centered cancer outcomes research. The study team met multiple times to summarize key informant interviews, integrate it with information from the literature, and organize the findings into categories and themes. These meetings were audiotaped to ensure capture of the entire discussion. Recommendations were collaboratively developed by the study team and reflect broad themes observed in the literature, results from the interviews, and specific issues or examples specified by multiple participants. Draft findings and recommendations were subsequently reviewed by select participants and other outcomes researchers to assure their accuracy and face validity. Lastly, the entire manuscript was reviewed by external experts participating in the AHRQ DEcIDE network.

RESULTS

Data Sources for Cancer CER

We identified 46 relevant datasets from our study sample of cancer outcomes researchers. Participants themselves expressed different opinions with regard to which data were important, adequate, or weak for cancer CER. This variability highlighted the lack of

[a]Associated with UNC Can-DEcIDE from: the University of North Carolina at Chapel Hill, Duke University, the Centers for Disease Control and Prevention, the Brigham and Women's Hospital, the University of Virginia, the Epidemiologic Research and Information Center at the Durham Veteran's Affairs Medical Center, the NC Central Cancer Registry, Blue Cross and Blue Shield of NC, Agency for Healthcare Research and Quality, and the National Cancer Institute.

standardized nomenclature associated with these data. Therefore, our first priority was to identify patterns with which we could organize existing datasets and broad themes. [10-29] Inspection of the datasets and their purposes revealed that a consistent nomenclature was needed before they could be easily organized to support CER. In response, an ontology was developed which divided the data into eight categories, including six "existing or fixed data" categories and two "hybrid" categories (Table 1).

Definitive empirical definitions for each category were difficult since they are not mutually exclusive. This was complicated by the fact that participants from different clinical and methodological specialties prioritized different characteristics of the datasets. Despite this, the study team reached consensus and unanimously agreed on the final ontology, which had face validity to internal and external reviewers, and provides useful classification and characterization of the datasets. Table 2 presents an illustrative sampling of what were perceived to be key datasets, assigned categories, and a summary of their strengths, limitations, and applicability for cancer CER, as-informed by the discussants and study team.

Barriers to the Conduct of Cancer CER

Several consistent themes emerged through the discussions and analysis: (1) There is a need for systematically identified, standardized measures to fill gaps and enhance data linkage and transferability. (2) Improvements in study design and population sampling are critical for CER studies to be meaningful. (3) Substantial issues exist regarding data ownership, access, governance, and cost. (4) Data security, privacy, and confidentiality remain paramount. (5) Broad multidisciplinary representation is needed to effectively address these CER data needs. These themes were consistent throughout the analysis and resonated with key informants and the study team.

DISCUSSION

We have developed a novel framework for organizing and characterizing cancer CER data together with relevant research needs. Based on the literature and key informant interviews, we propose a practical ontology regarding data resources and availability. The structure of this ontology was defined through a retrospective lens, by asking participants to nominate secondary sources of data that could be immediately leveraged or developed.

The retrospective lens provides a starting point from which a rational ontology can be developed. It allows us to define and characterize available data resources ready for cancer CER. And lastly, it provides a characterized delineation point, for transition from retrospective data models to prospective CER data models. Moving forward, we anticipate a transition to more frequent prospective and real-time data collection activities (electronic health records, continuously aggregating registries, rapid learning data systems). The ontology proposed from our study provides a foundational nomenclature from which to build future data resources. Increasingly clear throughout the fields of science and engineering is the need to organize and systematically structure data so that information can be maximally extracted to assist in prescribing the right treatment at the right time for a specific patient. Moreover, we need agreement and collaboration from the respective stakeholders of the multiple diverse systems for collecting data. This study highlights these realities and helps to point a practical way forward.

Our approach includes several limitations. Development of this ontology was challenged by a lack of mutual exclusivity among datasets and the diverse perspectives of participants. Our federally-funded study team was focused on describing datasets and CER opportunities with a government perspective. Our sampling was purposeful but not exhaustive; additional

cancer CER datasets have likely been missed, and the relative impact on the ontology is not clear.

Despite these challenges, this work provides a practical ontology that is adaptive and can be upgraded over time. It provides a template for understanding the strengths and limitations of current CER data resources, and formulating recommendations and guiding principles to advance cancer CER.

We present recommendations corresponding to the major themes identified in this study, with a goal of informing the evolution of the CER data framework, resolving data gaps, and ultimately establishing a national data infrastructure for cancer CER. Our focus was on existing secondary, observational data, though findings we present are also applicable to prospective data collection and future data resources.

1. There is a need for systematically identified, standardized measures to fill data gaps and enhance linkages and transferability

Inconsistent, incomplete measures and a lack of data standardization pose a substantial threat to improving public health through CER. Stakeholders (e.g., researchers, providers, payers) collect clinical, population, and health services data in numerous ways. Even within the research community, there are substantial differences of opinion on essential variables. This lack of consensus inhibits comparability across and within health datasets.

Recommendation 1a: Systematically identify necessary measures including uniform definitions and standardization of collection and coding—Intervention selection, exposure assignment, and outcome measures must be systematically identified and characterized. As a starting point, the study authors have recommended a framework for identifying measures across the cancer care continuum.[9] Standards for how measures are defined, collected, and coded must be developed and broadly applied, even for very basic measures such as race and ethnicity. This issue extends to algorithms for defining meaningful measures and cohorts, or deriving complex treatments or outcomes. Lack of global standardization inhibits data pooling, comparability among multiple sources, and generalizability of findings in the context of population heterogeneity.[30]

A multidisciplinary panel of CER researchers, stakeholders, and their partners is required to address this diversity of measures and lack of data standardization. A goal of such an effort should be identification of a minimum basic set of essential measures in all new data collection initiatives, including standardized data definitions.

Recommendation 1b: Develop and incorporate new measures and dataset crosswalks to address gaps among current data resources—Additional measures must be identified which incorporate advances in medicine and health sciences. A key example is the enhancement of our national cancer registries' collection of data on genetic markers. These tests, like the KRAS test, are increasingly able to provide predictive insight into intervention effectiveness for individual patients.[31-33] Because of the potentially rapid and inconsistent adoption of these markers, multi-concept coding systems are necessary to capture (1) if the test was used, (2) test results, and (3) test characteristics. In addition genetic markers, federal and other payers could consider standardization of clinical markers such as stage, grade, and performance status. The current utilization of ICD-9 and Healthcare Common Procedural Coding System (HCPCS) codes is insufficient in this cancer-specific context. Furthermore, investment in measurement and methods research could facilitate the development of 'crosswalks' between existing measures and instruments.[34, 35] This will enable the comparison of constructs between datasets and offer

potential mechanisms for combining existing data, or supplementing missing information.[36, 37]

Increasingly relevant for cancer CER are intermediate outcomes, including patient reported outcomes.[38-42] Historically, clinical research has focused on mortality, but through advances in cancer detection and treatment, patients are living longer and may not die from cancer. To enable better comparisons between treatments new measures are needed which go beyond life expectancy and better quantify side-effects, costs, and other trade-offs such as the probability of continuing to work or attending to family needs.[43, 44] Patient treatment decisions are increasingly likely to be informed by factors such as these. Systems to capture these measures must be better integrated into clinical care data, and embedded in future datasets.[30]

Recommendation 1c: Establish data architecture and systems standards for collecting and communicating these measures among health care delivery and financing organizations and researchers—To date, health care reform has focused on standards for patient care, transferability (Health Information Exchange [HIE]), and quality of care evaluations; however, CER also needs to be included as a priority component for improving health care. “Meaningful use” regulations offer significant incentives to standardize clinical data for transferability and interoperability, though these efforts are still nascent. Accordingly, CER stakeholder involvement is critical in the discussions between Centers for Medicare & Medicaid Services (CMS) and the Office of the National Coordinator for Health Information Technology (ONC), and must extend beyond meaningful use requirements for HIT development and requirements. NCI's cancer Biomedical Informatics Grid (caBIG) and Cancer Data Standards Registry and Repository (caDSR) have already developed an interoperable information technology (IT) infrastructure that offers standard rules, unified architecture, and common language to develop and use cancer research data.[30] It is vital that open-source, open-access tools such as these remain at the forefront of integrating with health care data coding such as Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), Logical Observation Identifiers Names and Codes (LOINC), or data interoperability such as HL7.

2. Improvements in study design and population sampling are critical for CER studies to be meaningful

Many of the problems with existing data sources cannot be solved through data standardization or sophisticated statistical methods. For example, a greater *quantity* of data will not necessarily make CER studies more generalizable or reproducible; rather, CER study design issues need to be better understood and overcome, resulting in better *quality* data.

While the focus of this work is not statistical methods or study design, data and study methodology are inexorably connected. Recognizing this, future studies need to prospectively apply more advanced data collection, better study designs, and sampling frameworks. At the same time, investments need to be made in ways to reduce bias through advanced statistical methods.[33] By funding research on study design issues in existing CER studies, we can develop better methods to apply toward future studies and data collection. It is also important to recognize that the advancement of complex methods requires consistency of measures and data interoperability described in the first recommendation.

Recommendation 2a: Develop methods to leverage existing data, overcome data limitations, and reduce bias—The majority of data currently used for CER is collected for non-research purposes and is non-experimental with regard to most CER

questions. Consequently, several significant sources of bias exist, some of which are correctable through advanced methods. Other sources of error are quantifiable, but cannot be adequately addressed. The development and application of better analytic methods can help overcome the design limitations of existing data. Propensity score matching and instrumental variable analysis are two important examples of statistical approaches that can capitalize on important data elements and advanced methods.

Many biases or data uncertainties can also be examined using specifically collected data or hybrid data sources. For example, linking administrative data to epidemiologic or clinical data (e.g., SEER-Medicare linked data[45]) creates powerful research resources that serve as models for other such efforts.[46] Other approaches include ancillary or validation studies collecting new data on a subgroup of the main population, or an external population, to supplement missing information or to extrapolate the distribution of an important variable into the study population.[47-50]

Recommendation 2b: Facilitate the conduct and completion of pragmatic trials for CER—Pragmatic trials can overcome many limitations of randomized clinical trials (namely, limited sample sizes and restrictive inclusion criteria). Pragmatic trials employ randomization but aim to make eligibility criteria and treatment decisions representative of "real world" settings.[4] They also collect information on a broader number of risks, determinants, health outcomes, and events, either directly or through the novel and efficient use of other data sources (e.g., claims/administrative data). As such, they can yield more generalizable findings. In addition to these benefits, increased funding for pragmatic trials could also help spur methods development on sampling and design issues commonly seen in traditional CER studies.[51]

3. Issues of data ownership, access, governance, and cost are substantial

There are many large data resources and innumerous small datasets relevant to cancer CER. However, there are significant barriers limiting their use including political obstacles, costs, and administrative burden associated with data access. [25] Important and timely data are often closely controlled by those who collect the data. Even data from federally funded studies may languish as the investigative team exhausts its "first right of publication." The potential benefits from additional data linkages are prevented by lack of access, cost, or tightly constrained data use agreements. For example, developing resources analogous to SEER-Medicare for the under-65 population is imminently feasible by linking registry data to private payer data. However, efforts to do so have commonly met with reluctance on the part of the payers and even registries. For these groups, research is not a primary priority, and the risks or "unknowns" are perceived to outweigh the prospective benefits.

Recommendation 3: Develop systems to facilitate timely data sharing for research supporting the public good—There are practical solutions to identify CER-relevant datasets and facilitate their acquisition.[52] This includes development of codified relationships among federal agencies, their contractors, and many data-holders.[25] For example, the individual SEER or NPCR registries could approve a single data acquisition process to be followed for all federally-contracted CER studies, which may relieve administrative burden. The National Cancer Institute's Central Institutional Review Board (IRB) may serve as a useful analog, as it was designed to relieve the work of the multitude of institutional IRBs.[53] However, it provides a cautionary tale, as the centralized IRB has been criticized for replacing rather than relieving the work needed to open a study.[54] Other examples include the broad DUAs between Medicare and important epidemiologic cohorts such as the Women's Health Initiative (WHI) study. Similar agreements could be developed for important cancer studies, making them more accessible to the research community.

Standardized relationships between state and federal agencies would help data-holders be reassured that their data will be used appropriately, while distilling data acquisition logistics to a formulaic process. These relationships would also help facilitate the timeliness of data for research and enable quick turn-around on important questions. Regarding access to costly or proprietary datasets, government stakeholders (e.g., AHRQ, NCI) may consider directly lending their weight to developing special agreements for select restricted or tightly-held datasets.

There may be utility in centrally-brokered and managed data subscriptions based on standing data use agreements. For example, states such as Maine and Oregon have implemented requirements that payers deposit "shadow claims" to public health agencies for purposes of quality improvement and informing policy decisions.[55] Formal mechanisms could be established to facilitate the updating and regular access to such data for CER.

4. Data security, privacy, and confidentiality remain paramount

While access to data must be improved, data security, privacy, and confidentiality are critical, and remain top concerns.[56] Moreover, there are multiple laws and regulations governing the maintenance, release, and use of many datasets, such as Medicare or Medicaid claims.

Recommendation 4: Develop systems to assure data security, privacy, and confidentiality with any enhanced access to data for CER—Two short-term practical opportunities warrant further exploration. First, at the state-level, health information exchange (HIE) is focusing on standardization of electronic health records and rules governing data transfer and use. It is prudent that the federal CER agenda be represented as new processes and regulations continue to be defined.[57-60] Second, developing a CER data security and utilization "accreditation" system may help assure compliance with regulations. This would ensure a baseline level of IT sophistication that facilitates data use while assuring data vendors that accredited research sites are top-tier, "safe" data custodians. Examining the Centers for Medicare and Medicaid (CMS) requirements of their quality improvement organizations (QIOs) [61] may be a first step to developing such accreditation systems.

5. Broad multidisciplinary representation is necessary to effectively address these CER data needs

The recommendations (from methods to policy) presented by this study's discussants are a microcosm of the larger CER discussion and highlight many differences in the cultures, values, terminology, measures, approaches, and priorities relevant for cancer CER, and are a microcosm of the larger CER discussion.

Recommendation 5a: A collaborative, multidisciplinary approach must be emphasized to successfully address data needs for CER—Multidisciplinary representation is necessary to adequately capture important differences in the cultures, values, and terminology surrounding perceptions of cancer CER and data issues.[26] Accordingly, a critical step will be identifying individuals who can represent their disciplines (and industries) to optimally advance the cancer CER discussion. To be successful, these individuals must not only be technical experts, they must also be mavens and translators who can bridge technical and disciplinary gaps to identify and achieve solutions.[62] Supporting the identification and ongoing communication of such a group will be important to drive CER data needs forward.

Recommendation 5b: CER stakeholders must be engaged and coordinated in the development of rules and standards to inform health reform—Beyond informing cancer CER and its data needs, it is important that a multidisciplinary advisory group be well represented in the context of health care policy reform. It will be vital to engage these diverse groups and address these issues in a timely and consistent manner – the recently established Patient Centered Outcomes Research Institute (PCORI) is the obvious choice to lead such an effort.[63] PCORI can identify members of the research community and partner with federal agencies. Both groups represent research need and interests and help define the future of CER in the context of health reform. Additionally, PCORI is well positioned to address other needs, such as maintaining a cancer CER data inventory, and perhaps similar registries for protocols including those with null results. The Registry of Patient Registries project is a promising project to begin to address this need.[64, 65]

Conclusions

By leveraging secondary data we can fill gaps and provide timely, valid, scientific knowledge to systematically conduct CER and improve cancer care and outcomes. However, substantial engagement is required from many organizations in order to address the issues outlined here. Multidisciplinary individuals within these organizations need to be identified who can help facilitate solutions in order for CER to reach its full potential.

The data ontology and recommendations we present provide guidance for critical discussions between multidisciplinary teams of cancer researchers, methods experts, and other stakeholders. They align with previous calls for infrastructure development to support cancer research and CER.[13, 15] Together they provide a template for systematically addressing cancer CER data needs. By understanding and overcoming weaknesses in current data, we can accelerate the pace of cancer CER, and ultimately enhance the adoption of CER findings to improve patient-centered care and outcomes.

Acknowledgments

We thank Timothy S. Carey, MD, MPH; and Janet K. Freburger, PhD; for their review and feedback which informed and strengthened this manuscript. We thank the anonymous reviewers from the Agency for Healthcare Resaerch and Quality (AHRQ)'s Effective Healthcare Program manuscript review system for their constructive suggestions and comments. This work was supported by funding from AHRQ through the Cancer DEcIDE Comparative Effectiveness Research Consortium, contract HHSA290-205-0040-I-TO4-WA5 – Data Committee for the DEciDE Cancer Consortium.

Funding Disclosures:

Dr. Abernethy has research funding from the US National Institutes of Health, US Agency for Healthcare

Research and Quality, Robert Wood Johnson Foundation, Pfizer, Eli Lilly, Bristol Meyers Squibb, Helsinn Therapeutics, Amgen, Kanglaite, Alexion, Biovex, DARA Therapeutics,Novartis, and Mi-Co; these funds are all distributed to Duke University Medical Center to support research. In the last 2 years she has had nominal consulting agreements (<$10,000) with Helsinn Therapeutics, Amgen, and Novartis.

References

1. Clancy CM, Slutsky JR. Commentary: a progress report on AHRQ's Effective Health Care Program. (AHRQ Update). Health Services Research. 2007; 42(5):xi(9).
2. Smith S. Preface. Medical Care. 2007; 45(10 Suppl 2):S1–S2. [PubMed: 18027399]
3. Congressional Budget Office. Research on the Comparative Effectiveness of Medical Treatments: Issues and Options for an Expanded Federal Role. Pub. No. 2975. Washington DC: 2007.
4. Maclure M. Explaining pragmatic trials to pragmatic policymakers. Journal of clinical epidemiology. 2009; 62(5):476–8. [PubMed: 19348972]

5. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA : the journal of the American Medical Association. 2003; 290(12):1624–32. [PubMed: 14506122]
6. Institute of Medicine. Initial National Priorities for Comparative Effectiveness Research. National Academies Press; Washington DC: 2009.
7. Sturmer T, Funk MJ, Poole C, Brookhart MA. Nonexperimental Comparative Effectiveness Research Using Linked Healthcare Databases. Epidemiology. 2011; 22(3):298–301. [PubMed: 21464649]
8. Sox HC. Defining comparative effectiveness research: the importance of getting it right. Med Care. 2010; 48(6 Suppl):S7–8. [PubMed: 20473202]
9. Carpenter WR, Meyer AM, Abernethy AP, Sturmer T, Kosorok MR. A framework for understanding cancer comparative effectiveness research data needs. Cancer Epidemiology, Biomarkers & Prevention Under Review.
10. Aday, L.; Begley, C.; Lairson, D.; Balkrishnan, R. Evaluating the Healthcare System: Effectiveness, Efficiency, and Equity. 3rd ed.. Health Administration Press; Chicago: 2004.
11. McDowell, I. Measuring Health. 3rd ed.. Oxford University Press; New York: 2006.
12. Lipscomb, J.; Gotay, C.; Snyder, C. Outcomes Assessment in Cancer: Measures, Methods, and Applications. Cambridge University Press; Cambridge: 2005.
13. National Cancer Policy Board of the Institue of Medicine. Ensuring Quality Cancer Care. National Academies Press; Washington DC: 1999.
14. National Cancer Policy Board of the Institue of Medicine. Assessing the Quality of Cancer Care: An Approach to Measurement in Georgia. National Academies Press; Washington DC: 2005.
15. National Cancer Policy Board of the Institue of Medicine. Enhancing Data Systems to Improve the Quality of Cancer Care. National Academies Press; Washington DC: 2000.
16. Zapka JG, Taplin SH, Solberg LI, Manos MM. A framework for improving the quality of cancer care: the case of breast and cervical cancer screening. Cancer Epidemiol Biomarkers Prev. 2003; 12(1):4–13. [PubMed: 12540497]
17. Shah, NR.; Stewart, WF. Clinical effectiveness: Leadership in comparative effectiveness and translational research.. Clin Med Res; the 15th Annual HMO Research Network Conference; Danville, Pennsylvania. April 26-29, 2009; 2010. p. 28-9.
18. Aiello Bowles EJ, Tuzzio L, Ritzwoller DP, et al. Accuracy and complexities of using automated clinical data for capturing chemotherapy administrations: implications for future research. Med Care. 2009; 47(10):1091–7. [PubMed: 19648826]
19. Sox HC, Greenfield S. Comparative effectiveness research: a report from the Institute of Medicine. Ann Intern Med. 2009; 151(3):203–5. [PubMed: 19567618]
20. Hoffman A, Pearson SD. ‘Marginal medicine’: targeting comparative effectiveness research to reduce waste. Health Aff (Millwood). 2009; 28(4):w710–8. [PubMed: 19556249]
21. Etheredge LM. Medicare's future: cancer care. Health Aff (Millwood). 2009; 28(1):148–59. [PubMed: 19124865]
22. Donabedian A. Evaluating the quality of medical care. 1966. Milbank Q. 2005; 83(4):691–729. [PubMed: 16279964]
23. Mandelblatt JS, Ganz PA, Kahn KL. Proposed agenda for the measurement of quality-of-care outcomes in oncology practice. J Clin Oncol. 1999; 17(8):2614–22. [PubMed: 10561329]
24. Brookhart MA, Sturmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. Med Care. 2010; 48(6 Suppl):S114–20. [PubMed: 20473199]
25. Bloomrosen M, Detmer D. Advancing the framework: use of health data--a report of a working conference of the American Medical Informatics Association. J Am Med Inform Assoc. 2008; 15(6):715–22. [PubMed: 18755988]
26. Bloomrosen M, Detmer DE. Informatics, evidence-based care, and research; implications for national policy: a report of an American Medical Informatics Association health policy conference. J Am Med Inform Assoc. 2010; 17(2):115–23. [PubMed: 20190052]

27. Manion FJ, Robbins RJ, Weems WA, Crowley RS. Security and privacy requirements for a multi-institutional cancer research data grid: an interview-based study. BMC Med Inform Decis Mak. 2009; 9:31. [PubMed: 19527521]
28. Safran C, Bloomrosen M, Hammond WE, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. J Am Med Inform Assoc. 2007; 14(1):1–9. [PubMed: 17077452]
29. Wallace PJ. Reshaping cancer learning through the use of health information technology. Health Aff (Millwood). 2007; 26(2):w169–77. [PubMed: 17259200]
30. Abernethy AP, Etheredge LM, Ganz PA, et al. Rapid-learning system for cancer care. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2010; 28(27): 4268–74. [PubMed: 20585094]
31. Lievre A, Bachet JB, Boige V, et al. KRAS mutations as an independent prognostic factor in patients with advanced colorectal cancer treated with cetuximab. J Clin Oncol. 2008; 26(3):374–9. [PubMed: 18202412]
32. Lievre A, Bachet JB, Le Corre D, et al. KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. Cancer Res. 2006; 66(8):3992–5. [PubMed: 16618717]
33. Sargent D. What constitutes reasonable evidence of efficacy and effectiveness to guide oncology treatment decisions? The oncologist. 2010; 15(Suppl 1):19–23. [PubMed: 20237213]
34. Cella D, Riley W, Stone A, et al. 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. 2010; 63(11):1179–94. [PubMed: 20685078]
35. Dorans NJ. Linking scores from multiple health outcome instruments. Quality of Life Research. 2007; 16:85–94. [PubMed: 17286198]
36. McHorney CA. Use of item response theory to link 3 modules of functional status items from the asset and health dynamics among the oldest old study. Archives of Physical Medicine and Rehabilitation. 2002; 83(3):383–94. [PubMed: 11887121]
37. McHorney CA, Cohen AS. Equating health status measures with item response theory illustrations with functional status items. Medical care. 2000; 38(9):43–59.
38. Abernethy AP, Zafar SY, Uronis H, et al. Validation of the Patient Care Monitor (Version 2.0): A Review of System Assessment Instrument for Cancer Patients. J Pain Symptom Manage. 2010
39. Abernethy AP, Ahmad A, Zafar SY, Wheeler JL, Reese JB, Lyerly HK. Electronic patient-reported data capture as a foundation of rapid learning cancer care. Med Care. 2010; 48(6 Suppl):S32–8. [PubMed: 20473201]
40. Clauser SB, Ganz PA, Lipscomb J, Reeve BB. Patient-Reported Outcomes Assessment in Cancer Trials: Evaluating and Enhancing the Payoff to Decision Making. J. Clin. Oncol. 2007; 25(32): 5049–50. [PubMed: 17991919]
41. Lipscomb J, Reeve BB, Clauser SB, et al. Patient-Reported Outcomes Assessment in Cancer Trials: Taking Stock, Moving Forward. J. Clin. Oncol. 2007; 25(32):5133–40. [PubMed: 17991933]
42. Lipscomb J, Gotay CC, Snyder CE. Patient-reported outcomes in cancer: A review of recent research and policy initiatives. Ca-a Cancer Journal for Clinicians. 2007; 57(5):278–300. [PubMed: 17855485]
43. Carpenter WR, Peppercorn J. Beyond toxicity: the challenge and importance of understanding the full impact of treatment decisions. Cancer. 2009; 115(12):2598–601. [PubMed: 19365843]
44. Hassett MJ, O'Malley AJ, Keating NL. Factors influencing changes in employment among women with newly diagnosed breast cancer. Cancer. 2009; 115(12):2775–82. [PubMed: 19365847]
45. Warren JL, Klabunde CN, Schrag D, Bach PB, Riley GF. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. Med Care. 2002; 40(8 Suppl):IV–3-18.
46. Jutte DP, Roos LL, Brownell MD. Administrative Record Linkage as a Tool for Public Health Research. Annual Review of Public Health. 2011; 32:91–108. Vol 32.
47. Goldberg RM, Sargent DJ, Morton RF, et al. NCCTG Study N9741: leveraging learning from an NCI Cooperative Group phase III trial. Oncologist. 2009; 14(10):970–8. [PubMed: 19828593]

48. Sturmer T, Glynn RJ, Rothman KJ, Avorn J, Schneeweiss S. Adjustments for unmeasured confounders in pharmacoepidemiologic database studies using external information. Med Care. 2007; 45(10 Supl 2):S158–65. [PubMed: 17909375]

49. Sturmer T, Schneeweiss S, Brookhart MA, Rothman KJ, Avorn J, Glynn RJ. Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: nonsteroidal antiinflammatory drugs and short-term mortality in the elderly. Am J Epidemiol. 2005; 161(9): 891–8. [PubMed: 15840622]

50. Sturmer T, Schneeweiss S, Rothman KJ, Avorn J, Glynn RJ. Performance of propensity score calibration--a simulation study. Am J Epidemiol. 2007; 165(10):1110–8. [PubMed: 17395595]

51. Holve, E.; Pittman, P. A First Look at the Volume and Cost of Comparative Effectiveness Research in the United States. Academy Health; Washington D.C.: 2009.

52. Conway PH, VanLare JM. Improving Access to Health Care Data The Open Government Strategy. Jama-Journal of the American Medical Association. 2010; 304(9):1007–08.

53. Wagner TH, Murray C, Goldberg J, Adler JM, Abrams J. Costs and benefits of the national cancer institute central institutional review board. J Clin Oncol. 2010; 28(4):662–6. [PubMed: 19841324]

54. Dilts DM, Sandler AB, Cheng SK, et al. Steps and time to process clinical trials at the Cancer Therapy Evaluation Program. J Clin Oncol. 2009; 27(11):1761–6. [PubMed: 19255315]

55. Office for Oregon Health Policy & Research. [April 12, 2010] Policy Brief: All-Payer, All-Claims Data Base. Available from URL: http://www.oregon.gov/OHPPR/HFB/docs/2009_Legislature_Presentations/Policy_Briefs/PolicyBrief_AllPayerAllClaimsDatabase_4.30.09.pdf?ga=t

56. Safran C, Bloomrosen M, Hammond WE, et al. Toward a national framework for the secondary use of health data: An American Medical Informatics Association white paper. Journal of the American Medical Informatics Association. 2007; 14(1):1–9. [PubMed: 17077452]

57. U.S. Department of Health and Human Services: Office of the National Coordinator for Health Information Technology. Nationwide Privacy and Security Framework For Electronic Exchange of Individually Identifiable Health Information December 15, 2008. Available from URL: http://bit.ly/2DVP5p

58. Organization for Economic Co-operation and Development. [June 6, 2010] OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data. Available from URL: http://www.oecd.org/document/18/0,3343,en_2649_34255_1815186_1_1_1_1,00.html

59. U.S. Department of Health and Human Services: Health Information Technology. [April 16, 2010] Meaningful Use. Available from URL: http://healthit.hhs.gov/portal/server.pt?open=512&objID=1325&parentname=CommunityPage&parentid=1&mode=2

60. North Carolina Healthcare Information and Communication Alliance. [June 2, 2010] North Carolina Health Information Exchange Council (NC HIE Council). Available from URL: http://www.nchica.org/GetInvolved/NCHIE/intro.htm

61. U.S. Department of Health and Human Services: Medicare.gov.. [May 28, 2010] What is a Medicare Quality Improvement Organization (QIO)?. Available from URL: https://questions.medicare.gov/app/answers/detail/a_id/1943/related/1

62. Gladwell, M. The Tipping Point: How Little Things Can Make a Big Difference. Little, Brown & Company; Boston: 2000.

63. Clancy C, Collins FS. Patient-centered outcomes research institute: the intersection of science and health care. Sci Transl Med. 2010; 2(37):37cm18.

64. Outcome Sciences. [January 27, 2011] Registry of Patient Registries (RoPR). Available from URL: http://www.outcome.com/ropr.htm

65. U.S. Agency for Healthcare Research and Quality. [February 6, 2011] 9. Registry of Patient Registries. Available from URL: http://www.ahrq.gov/fund/recoveryawards/osawinfra.htm

66. [November 11, 2010] PatientsLikeMe I. Available from URL: http://www.patientslikeme.com/

67. Sturmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. Am J Epidemiol. 2005; 162(3):279–89. [PubMed: 15987725]

Table 1

Data ontology and definitions.

Existing and Fixed data	**Definition**
Experimental studies data	Primary data collected for the purpose of studying the safety and efficacy of health and medical interventions. Significant variation exists between types of studies with regards to utility for CER. For example, phase III-IV randomized trials are more limited for CER versus pragmatic trials or large epidemiologic trials (e.g., WHS, WHI, PHS).
Non-experimental studies data	Data collected by public health researchers to identify patterns and determinants of diseases and outcomes within a population (e.g., NHS, HPFS, ARIC).
Registry data	Data collected by public health or other clinical health institutions to evaluate disease incidence, morbidity, and outcomes for a population defined by a specific condition or exposure, including interventions (e.g., SEER)
Administrative / claims data	Data collected for the business or programmatic purposes of documenting, delivering, or paying for health care, including insurance companies, health systems, or government entities. May be for the purpose of organizing, tracking and defining patient health and interactions with the healthcare system (e.g., Medicare, Medicaid, MarketScan).
Electronic Health Records	Data collected at point of care to support clinical care delivery, management, and decision making. These data are stored/managed through specific computer-based software and information systems (e.g., HMO-network).
Other data	Various, yet untested for CER; examples include syndromic surveillance and or pharmacy purchase/market data.
Hybrid data	
Linked clinical and claims data	Datasets created by the linkage between two unique data sources (often from the above categories) collected by different entities and for different purposes (e.g., SEER-Medicare)
Validation study data	Data collected or obtained for the purposes of overcoming limitations from existing/secondary data sources (e.g., MCBS, POC)

Abbreviations: WHS: Women's Health Study; PHS: Physicians' Health Study; NHS: Nurses Health Study; HPFS: Health Professionals Follow-up Study; ARIC: Atherosclerosis Risk in Communities Study; MCBS: Medicare Current Beneficiary Survey; POC: Patterns of Care Study; etc.

Table 2

Existing and secondary data sources: Illustrative examples, strengths, limitations and applicability/utility to CER.

Examples	Strengths	Limitations	Applicability/Utility to Cancer CER
Existing and Fixed Data Sources			
1) Experimental studies (trials) data (e.g., clinical trials, pragmatic trials):			
■ ACCENT/ PS2 ■ Women's Health Initiative (WHI) ■ Physicians' Health Study(PHS) ■ Women's Health Study (WHS)	■ Detailed and unbiased information on treatment, and important clinical covariates ■ Enormous breadth and diversity of data (across 12 NCI cooperative groups)	■Limited generalizability ■ Expensive to conduct, requires lengthy follow-up for many outcomes ■ Limited sample sizes ■ Highly specific (i.e. usually single treatment/intervention) ■ Limited in essential/important covariates	■Utility for CER depends on type of experimental study ■ Broadly defined (or population-based) trials can be useful for CER; but require extensive inclusion of covariates outside of the main trial aims ■ Secondary use of experimental studies for CER could be improved through investments in 1) Pragmatic clinical trials and 2) methods / design development
2) Non-experimental (observational) studies data			
■North Carolina - Lousiana Prostate Cancer Project ■ Cancer Care Outcomes Research and Surveillance Consortium (CanCORs) ■ Health Professionals Follow-Up Study(HPFS) ■ Nurses Health Study (NHS) ■ American Cancer Society Cohort	■Extensive data on diagnosis, procedures and outcomes ■ Rich in covariates (risk factors, important confounders) ■ Often include patient medical records ■ Can be population-based	■Expensive to develop and maintain ■ Logistics of study development limit data availability and addition of new hypotheses ■ Several biases may exist: selection; information; recall; and response ■ Unclear event temporality between data collection waves ■ Limited in scope, statistical power beyond initial study aims ■ Proprietary data requiring extensive protocols, procedures	■ Can be leveraged for comparative effectiveness depending on data quality and extent of biases ■ Utility for CER also dependent on study design, quality/completeness of measures and broad inclusion covariates ■ Can be strengthened through potential data linkages to claims or EHR data which can augment or off-set biases/limitations (can provide temporality of events, verification of treatment/outcomes, etc.)
3) *Registry data*			
■Surveillance Epidemiology and End Results (SEER) ■ National Program of Cancer Registries (NPCR) ■ National Cancer Data Base (NCDB) ■ National Oncologic Positron Emission Tomography (NOPR)	■Rich disease information ■ Clinical information at point of care or diagnosis ■ Simultaneously collected with diagnosis and treatment ■ Opportunity for recruitment into cohorts or trials ■ Can link with administrative data	■Potential sampling biases (selection, inclusion, etc.) ■ Questionable generalizability ■ Primarily limited to first occurrence of event or disease and limited inclusion of covariates ■ Unknown response, toxicity, patient reported outcomes ■ Challenging for longitudinal data capture ■ Sparse patient identifiers ■ Challenging for selecting controls / comparator populations	■Do not provide enough complete data for rigorous CER ■ Linkages to additional data are necessary to provide missing information ■ Dearth of literature on solutions/ methods for inherent biases, interoperable study design, and evaluation/application of comparator populations
4) Administrative and claims data			
■Most health insurance programs: Medicare; Medicaid; Blue Cross / Blue Shield, etc. ■ Medstat / Marketscan ■ United Health	■Represents large proportion of US population ■ Rich patient-level data: demographics, procedures, treatments ■ Includes temporality of events ■ Some include organizational/ provider characteristics ■ Most have unique identifiers enabling linkage to other data	■Design/structure often impacts data sensitivity/specificity ■ Missing important clinical etiologic information ■ Includes date or type of testing procedures, but no results (e.g., pathology, tumor response, genetics, vital stats, etc.) ■ High patient turn-over ■ Complicated data structure requires significant learning-curve and programming resources ■ Burdensome and prohibitive data use agreements ■ Expensive to obtain	■Missing key CER components including vital tumor and disease information ■ Linkages can supplement missing information – but costs and/or DUA's often inhibit additional linkages ■ Utility for CER would be greatly improved through institutional and governmental policies which overcome limitations (i.e., funding, training, collaboration)

Examples	Strengths	Limitations	Applicability/Utility to Cancer CER
		■ Untimely data releases – significant time lags	
5) Electronic health records			
■Health care systems: Veterans Administration (VA); HMO-network; Kaiser; Mayo; Geisinger; US Oncology ; UK General Practitioners Research Database (GPRD) ■Large vendors: GE Health; Allscripts/Misys; Epic; McKesson; NextGen	■Includes multiple data components (practice management, electronic patient record, patient portal) ■Fully integrated EHR's provide clinical information, claims, tumor specifics, longitudinal follow-up, objectively measured events ■Allows for studies of toxicity, quality of life, natural history	■Populations are not generalizable ■Lack of standardization of patient information and clinical measures between systems (technology, data structure, and coding) ■Missing or insufficient data elements necessary for CER ■Imperfect record keeping/follow up - Patients not consistently maintained within a single system/ EHR ■Enormous expense to obtain data from private sector/vendors	■Currently there is limited utility for EHR data from private vendors ■However examples from VA and universal/national systems (UK, Canada), exemplify potential of EHR sources ■Future utility dependent on: standardization of measures *and* data systems/interoperability; standard linkage variables; public *and* private institutional data governance and stewardship
6) Other Data			
■Genetic and genomic data ■Geospatial data ■Environmental monitoring data ■Over the counter drug purchasing ■Health seeking on internet ■Patient-networking sites, [66] ■Syndromic surveillance	■Data at both patient and ecological level ■Information on behavioral and environmental risks ■Can provide information on disease determinants ■Self-reported experiences, exposures, outcomes	■ Unclear how to identify, define and utilize these data	■ Utility to CER dependent on integration into other data, specifically clinical care data
Hybrid Data Sources			
7) Linked clinical and claims data			
■SEER-Medicare ■State Cancer Registry – Medicare/Medicaid ■WHI-Medicare	■Includes clinical and health services data ■Provides temporality of events ■Large population samples; ability to study rare events/ treatments ■Provides access to controls or comparison populations ■Allows for adjudication/ validation of events (i.e., self-reported) ■Can detecting recurrence	■Missing information (eg, HMO or supplemental insurance); often highly specific populations (>65, disabled, etc) ■Non-covered services are excluded (e.g., prescription drugs, long-term care, free screenings) ■Missing vital clinical information (tumor response) ■Treatment rationale and test results are unknown ■Complicated algorithms needed to characterize treatment ■Large, complex data require advanced training/experience ■Delay in research access	■Powerful for CER studies because of large, generalizable populations ■Large number of covariates and clinical information ■Lengthy follow-up available including information on temporality of treatment and events ■Could be strengthened by linkages to laboratory and clinical results
8) Validation study data			
■Internal validation studies[49, 67] ■External validation studies[47]	■Rich disease information ■Used to minimize limitations of other data ■Can give estimates of associations not discernable within data	■Lack of validated studies exist for CER ■Methodologic limitations and lack of model transportability to CER	■To be useful for CER an investment in methodologic work is required -- similar to P01 CA 142538 “Statistical Methods for ■ Cancer Clinical Trials” (PI, Kosorok) ■Validation studies could lead to immediate return of investment with regard to leveraging existing data for CER