Oncoshare: Lessons Learned from Building an Integrated Multi-institutional Database for Comparative Effectiveness Research

Susan C Weber[1], Tina Seto[1], Cliff Olson[2], Pragati Kenkare[2], Allison W. Kurian[3,4], Amar K. Das[4]

[1]Center for Clinical Informatics, Stanford University
[2]Palo Alto Medical Foundation (PAMF) Research Institute
[3]Department of Health Research & Policy, Stanford University
[4]Department of Medicine, Stanford University

Abstract

Comparative effectiveness research (CER) using observational data requires informatics methods for the extraction, standardization, sharing, and integration of data derived from a variety of electronic sources. In the Oncoshare project, we have developed such methods as part of a collaborative multi-institutional CER study of patterns, predictors, and outcome of breast cancer care. In this paper, we present an evaluation of the approaches we undertook and the lessons we learned in building and validating the Oncoshare data resource. Specifically, we determined that 1) the state or regional cancer registry makes the most efficient starting point for determining inclusion of subjects; 2) the data dictionary should be based on existing registry standards, such as Surveillance, Epidemiology and End Results (SEER), when applicable; 3) the Social Security Administration Death Master File (SSA DMF), rather than clinical resources, provides standardized ascertainment of mortality outcomes; and 4) CER database development efforts, despite the immediate availability of electronic data, may take as long as two years to produce validated, reliable data for research. Through our efforts using these methods, Oncoshare integrates complex, longitudinal data from multiple electronic medical records and registries and provides a rich, validated resource for research on oncology care.

Introduction

Comparative effectiveness research (CER), a well-established framework for evaluating health outcomes, became a national priority as part of recent health care reform legislation [1]. The use of observational data, retrospectively collected from existing electronic resources, is one of the primary approaches to undertaking CER [2]. Prior work on building such CER data resources, such as those undertaken in the practice-based DARTNet [3] and at the Veterans Administration Healthcare System [4], indicates the need to integrate and standardize data from multiple resources. In this paper, we present our efforts, as part of the Oncoshare project, to address the specific challenges of extracting and linking patient-specific data from multiple institutional databases, standardizing a database schema and terminology for data integration, and prioritizing data sources in deriving outcome measures.

Oncoshare is a collaborative research project that draws on expertise in medical informatics, clinical oncology, health policy, and epidemiologic research from the academic medical center of Stanford University, the research institute of the community-based Palo Alto Medical Foundation (PAMF), and the research group at the regional Cancer Prevention Institute of California (CPIC). Launched in 2009, the goal of the Oncoshare project is to develop a shared database for translational research and outcomes analysis of data on women treated for breast cancer across academic and community settings. Retrospective electronic data on women treated for breast cancer were available from these institutions for 10 years spanning 2000 to 2009.

Prior to the development of the Oncoshare data resource, the project team selected the following questions to pursue comparative effectiveness research:

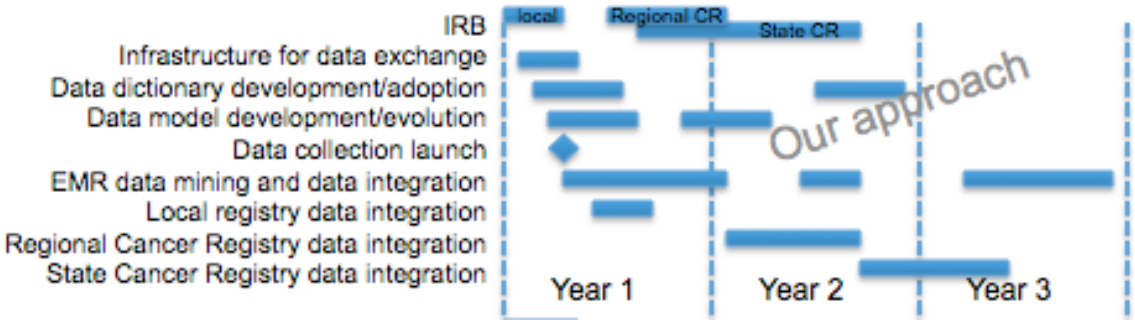

Figure 1: Oncoshare project task Gantt chart (approximate)

- How do treatment patterns deviate from current practice guidelines and across clinical settings?
- How do treatment deviations relate to patient outcomes?
- What biomarkers, clinical and demographic factors relate to patient outcomes?
- Has the clinical application of emerging diagnostic tests, such as tumor genomic assays and magnetic resonance imaging, affected patient and physician choices about adjuvant therapies for breast cancer?
- Is the use of emerging diagnostic tests improving clinical outcomes?

Methods

The timeline followed by the project is depicted in Figure 1, and indicates the specific efforts to address human subjects issues for retrospective extraction, validation and integration of electronic data into a CER database; to build an infrastructure for data exchange between institutions; to develop standard data dictionary and data model for the database; to determine inclusion of subjects into the database; and to integrate registry data into the database. We present our methods for each approach in the following section.

Addressing Human Subjects Issues: The initial task was to set up the necessary approvals by the Institutional Review Boards (IRB) and Privacy Officers at each institution. Each institution initially wrote two separate IRB protocols: 1) one to cover the activity of assembling the de-identified data set and accurately linking patients held in common while minimizing the sharing of protected health information (PHI) as described in Weber et al [5], and 2) a second to cover the researchers' use of the assembled joint data set, the Oncoshare database proper. Dataset assembly work in local staging databases was covered by an expedited IRB protocol with waiver of consent (which was feasible because this data set did not contain pediatric patients), and use of the Oncoshare dataset for research purposes was covered by a determination of non-human subjects research obtained from the IRB. The resulting system architecture is depicted in Figure 2. Future research projects using the Oncoshare database will require separate IRB approvals.

Building an Infrastructure for Exchanging Data and Linking Patients: While waiting for local IRB approval, we proceeded with building the infrastructure required. We did not wish to designate one site as the primary holder of the data, but rather wished to share the data between our two institutions such that both sides would always have ready access to the conjoined data set. We considered a grid-like architecture such as caGrid [6] but opted instead for a non-scalable point-to-point data sharing solution to avoid taking time and resources away from the primary task of data harmonization and integration. A Virtual Private Network (VPN) connection was accordingly established between our two secure data centers so that data could be directly transferred from one database to the other. A grid architecture could be implemented in the future if the need arose to add more nodes, to support faster refresh cycles, or on the fly data requests.

The data exchanged in this fashion were limited to the de-identified data; the only identified data exchanged were in support of verifying the identities of patients held in common, as covered by the dataset-assembly protocols

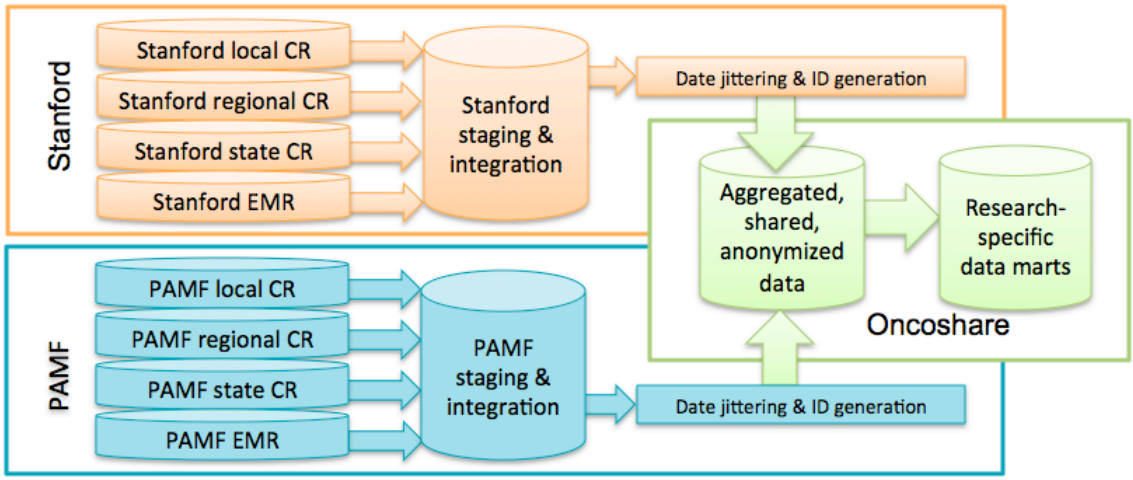

Figure 2: System architecture. Data staging activities were covered under IRB protocols, and the resulting de-identified Oncoshare database covered under a determination of non-human subjects review

approved by the IRBs and Privacy Officers. While we consider our data set to be de-identified, we also put into place the necessary data use agreements required when sharing limited data sets as per the Privacy Rule of the Health Information Portability and Accountability Act (HIPAA).

One challenge we faced was how to accurately identify patients held in common between the two sites while exchanging the minimum necessary PHI, according to requirements of HIPAA. We determined that each side should generate a study identification code based on patient name and date of birth, which would have a high likelihood of key collision in cases of patients seen at both sites. Exact matches on full name and date of birth would significantly under-count the patients held in common, due to spelling inconsistences and even name change, so we experimented with using a short prefix of the patient names and found that date of birth plus the first two letters of the patient's first and last name had the desired characteristics, as described in detail in Weber et al [5].

Since dates are considered PHI as per HIPAA, we needed to specify temporal information without disclosing the actual date on which the provider-patient interaction occurred. We considered various approaches [7] and settled on date offsets. While date shifting may not be legally tested under the safe harbor provisions, the use of date shifting is a published method of meeting the safe harbor requirements [8,9,10]. A random offset between -31 and +31 was generated for each patient, based on the patient's unique identifier as follows:

If (Hash_Code is even), then Offset = (Hash_Code mod 31) + 1;
If (Hash_Code is odd), then Offset = 0 - ((Hash_Code mod 31) + 1));

A prime number of 31, rather than 30, worked better with the modulus function to create a more even distribution of random offsets. The resulting number was used to systematically offset every date of service for the patient, so that the precise temporal sequence of service was preserved. Offsetting by a month in either direction also relatively preserved any seasonal significance. The algorithm for generating the hash code from the patient identifier was shared only between IRB-approved data staging staff, and the offset itself is sequestered in the staging environment along with all the true dates.

In an effort to preserve anonymity of healthcare providers we generated per-provider codes consisting of a short string hash of their California medical license number. The provider hash id was augmented with the provider's specialty to facilitate categorization of providers, such as 065483A-MEDICAL ONCOLOGY.

Developing a Standard Data Dictionary and Data Model: An initial investigative aim of the project was to apply standard terminologies to encode data on patient-provider interactions. To that end we developed a data dictionary based on National Cancer Institute (NCI) codes for the data elements that we planned to obtain from our electronic medical records (EMR). We also leveraged prior work mapping medication names to RxNorm codes [11] for

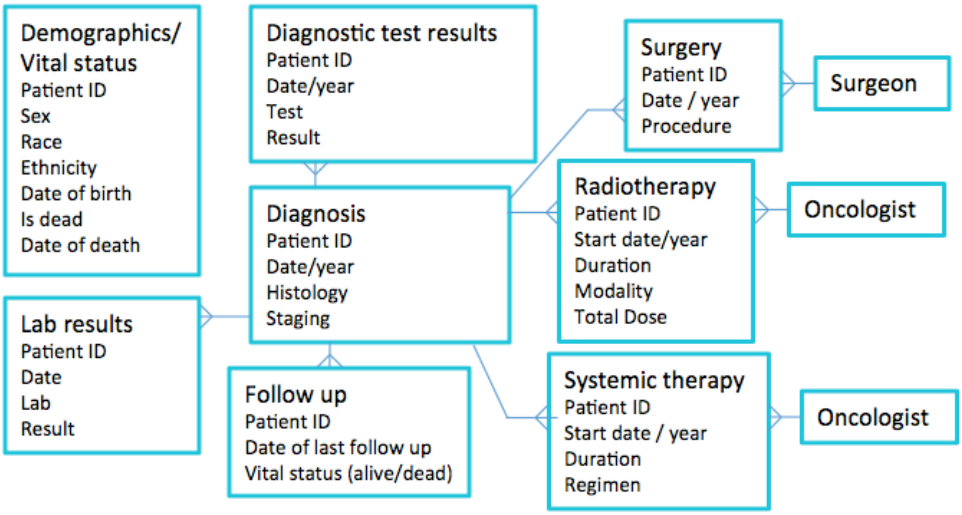

Figure 3: initial data model. All dates, including date of birth and death, are systematically 'jittered' by a per-patient offset between -31 and 31 days

consistency in coding across both sites. The data model we proposed initially is depicted in Figure 3. We envisioned that this data model would provide an EMR-independent, rich representation of a patient's journey through diagnosis, treatment and recovery.

Determinining Inclusion Criteria for Patients: After we obtained IRB approval, we began collecting and exchanging data. The first challenge was to determine a precise definition of which patients should be included in the Oncoshare database. Our initial inclusion criterion was simply stated as patients treated at Stanford (or PAMF) for breast cancer. However, determining whether a patient seen in clinic was actually being treated for breast cancer proved surprisingly complicated. We performed extensive chart review on hundreds of patients to clarify our understanding of this issue. In some cases, we found that ICD-9 codes for breast cancer were used incorrectly for patients undergoing prophylactic treatments such as mastectomy or tamoxifen, when these patients never actually had breast cancer. In other cases, we found that breast cancer survivors were treated with cancer therapies (such as chemotherapy or radiation) for an unrelated subsequent tumor. This difficulty in accurately characterising our cohort exclusively through use of data available in the EMR led us to contact our hospitals' tumor registries. Upon examination of the hospital cancer registries databases, we determined that they would make a more robust starting point than the EMR, consisting as they do of concise, semi-structured, hand-curated records gleaned by tumor registrars who abstract clinical narratives from the EMR.

Integrating Registry Data into the Database: In an effort to gain as much information about the patient as possible (especially staging data), we approached our regional cancer registry (CPIC) and added their investigators to the data assembly IRB protocol. The process of obtaining permission from the California state IRB for use of regional registry data took 6 months, followed by an additional 6 months to gain access to state-wide registry data. The state-wide data proved the more useful starting point, since it contained the accrued record of a patient's treatment throughout centers across the state of California, not just from the local San Francisco Bay Area. We found that state-wide registry data contained a more consistent and complete historical record of disease pathology (tumor TNM staging) and of major treatment modalities than we had been able to compile by mining our EMR.

Results

The clinical informatics team met weekly over a two-year period to undertake the steps presented in the methods section. Through an iterative development process, we created the final data model for Oncoshare that is depicted in Figure 4. Given the multi-disciplinary, often multi-institutional nature of breast cancer care, this simpler model proved more feasible than the one we had originally proposed (Figure 3). This model was driven by the need to find a common ground between the different EHR systems implemented at the two sites as well as historical EHR

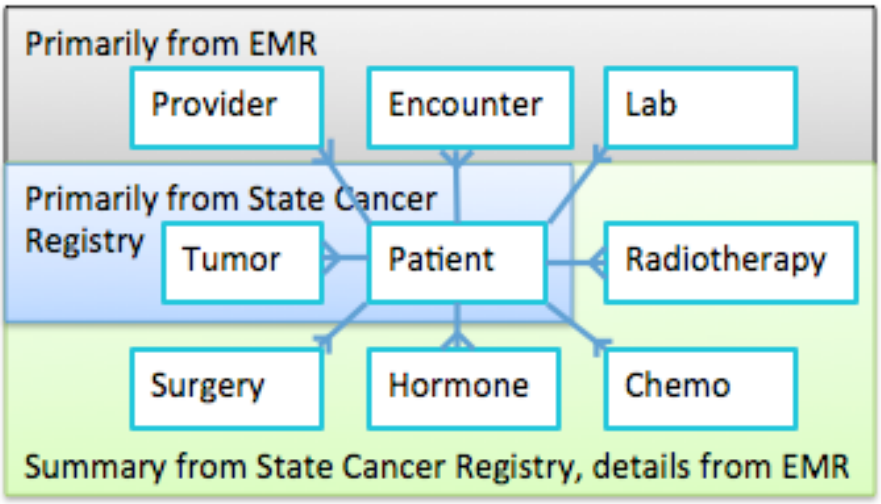

Figure 4: final data model and primary sources for each data type. In this model relationships between tumor and treatment must be inferred by temporal proximity

migrations at each site. These system differences were identified, reviewed, and analyzed constantly during these weekly meetings. The final data model reflects the compromises necessary to integrate the data from multiple systems across entities and time.

Our initial dataset, compiled from our respective institutional EMRs from 2000 to 2009 inclusive, contained 8390 patients seen at PAMF and 11,010 at Stanford, 2137 of which were held in common (consisting of approximately 25% of the PAMF patients and 20% of the Stanford patients).

Clinical encounter data drawn from the EMR consisted of billing codes (the code, source vocabulary and descriptive text), patient, provider and date. This table was richly populated, averaging 80 encounter codes per patient, with all columns complete.

Tumor data drawn from the EMR and cancer registry consisted of patient identification code, date of diagnosis, stage and expression of clinically important tumor markers (estrogen receptor (ER), progesterone receptor (PR) and HER2/neu), histology, and pathologic Tumor, Node, Metastasis (TNM) staging. This table was sparsely populated, with TNM scores, staging and tumor markers available on only 50% of the patients originally identified by billing diagnosis codes alone. Due to significant rates of missing data, we defined an analytical cohort (N=12,116) of patients having adequate information for characterization of their breast cancer, including stage, tumor markers, and evidence of some treatment or diagnostic information.

Surgery data, while potentially available in the EMR, was most reliably extracted from the cancer registry, particularly for patients with complete staging, histology and tumor marker information reported by the registry. We were confident in interpreting a missing report in the state-wide cancer registry as a true absence of surgery, whereas a missing report in the EMR might reflect a billing error or performance of the procedure at a different hospital.

Given recent reports that SEER may under-ascertain specific treatment modalities, particularly radiation therapy [12], we used EMR data to supplement the registry summary of treatment for each complex major modality, namely systemic therapy and radiotherapy. Efforts to add specific details of chemotherapy regimens, such as drug combinations, doses and intervals, proved challenging given the evolution of EMR-based drug ordering over the last decade. We anticipate that prospective data capture of chemotherapy regimens will prove more straightforward, with the increasing use of chemotherapy-specific electronic ordering programs such as Beacon. A major contribution from the EMR was the addition of billing codes for emerging diagnostic interventions including imaging strategies, genetic and tumor genomic tests; this information was not available through the cancer registry.

We obtained survival data from the state cancer registry according to their reported algorithm interrogating multiple national databases for last date of follow up [13] and by querying the vital status field. We also integrated data from the Social Security Administration Death Master File (SSA DMF) [14]. We used a consistent algorithm for patients seen at both institutions, to minimize any bias in death ascertainment. Since only 85% of our patients have provided us with a seemingly valid SSN as part of our normal registration process [5], the heuristic used to match EMR patients to the SSA DMF is as follows:

1. Match with the patient's full SSN and the month/year of the birthdate (YYYYMM)
2. If not found, match with the patient's exact last name, exact first name, full birthdate (YYYYMMDD) and the last 4 digits of the SSN.

We are currently using the Oncoshare database to prepare a manuscript that will report on the patterns and outcomes of breast cancer care across these community and academic health systems over the last decade. Additionally, we are now collaborating with patient advocates to develop questionnaires for collection of patient-reported information on care preferences, symptoms, and outcomes, which we will integrate into Oncoshare [15, 16]. We plan on using and expanding this resource for other projects over time, e.g. adding genetic testing results and investigating their implications in treatment and outcomes.

Discussion

In undertaking an iterative process of developing a validated data resource for CER in oncology, we were able to evaluate which of our methods were most efficient in addressing the informatics challenges of selecting, extracting, standardizing, sharing, and integrating data from existing electronic resources. As a result of our efforts, we recommend the following sequence of tasks to develop a retrospective, multi-center cancer research database (Figure 5):

1. Apply for IRB approval at each participating site, and apply for the relevant extract from the state cancer registry at the project outset. We found it necessary to identify breast cancer patients locally and send our state cancer registry their names and dates of birth, in order to receive a report on a subset of patients that were matched to the state files.

2. Build the local data-sharing infrastructure. IRB approval is not required to implement machinery for data exchange (although approval is required to use such machinery), so this infrastructure can be built while awaiting approval; the patient identifier generation and date jittering machinery can also be built and tested during this interval.

3. Review the SEER data dictionary in order to specify which variables to be included in the cancer registry report. Asking for "all" variables may provide many that will not be used, and may slow the data transfer process. We recommend using applicable codes in the SEER data dictionary in the shared data set, and adopting new codes only for the local extensions to the data. Before adopting new codes, check with the statisticians who will be performing the research study analyses: statisticians may prefer all-numeric codes, which rules out the choice of NCI codes.

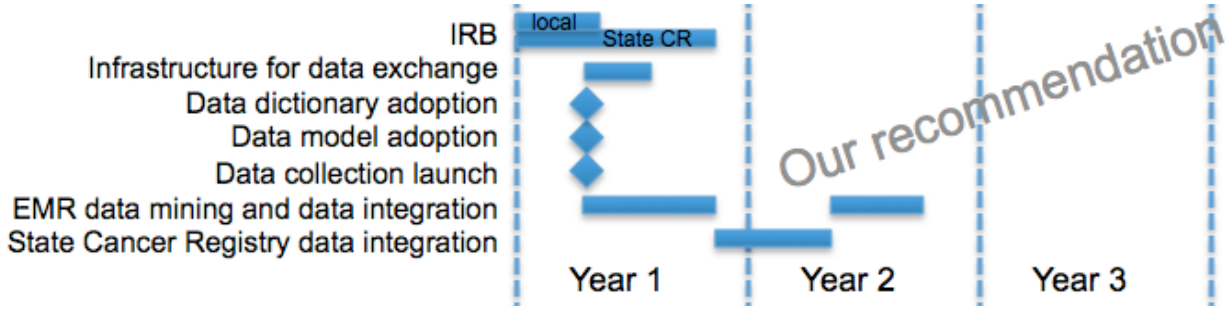

Figure 5: recommended project timeline for building a retrospective cancer database

4. Create tables that capture patient demographics, patient-provider encounters, tumor details, and treatment details. We recommend one table for each major treatment modality, laboratory results, and provider names and affiliations (if provider identity is a component of your research). Patients with single cancer diagnoses present a clear picture in this model, with temporal proximity a strong indication of causality; patients with co-occurring cancers of different organ sites prove more challenging, due to the difficulty of accurately associating a treatment to a tumor when there is more than one (and the decision may be made to exclude such patients, who are relatively rare, from statistical analyses). We recommend prospective manual data capture to build a dataset with a complex underlying data model.

5. Following IRB approval for sharing data, exchange patient information and validate the list of patients held in common. At this stage, it is also recommended to share local EMR encounter data, including provider details, laboratory test results, costs, and any other relevant (and permissible) data by way of validating the data exchange machinery and processes. It is likely that the final study cohort will remain in flux at this point. We recommend obtaining and integrating the Social Security Administration Death Master File, if survival is a desired study endpoint.

6. After receipt, processing and sharing of all participating institutions' respective state cancer registry datasets, the study cohort can be finalized as those patients included in the state report. We recommend mining institutional EMRs for clinical information including:
    a. Drug administration detail: actual dose and dates of administration
    b. Drug treatment protocol metadata: combination chemotherapy regimen, number of cycles, interval and supportive agents
    c. Diagnostic tests absent from cancer registries, such as imaging technologies (magnetic resonance imaging, positron emission tomography) and genetic tests (such as BRCA1/2 mutation testing), which may associate with care and outcomes
    d. Tumor-specific diagnostics, such as expression of ER, PR, and HER2/neu, or tumor genomic profiling with the 21-gene Recurrence Score [17]) for breast cancer
    e. Clinical trial information, when available
    f. Radiotherapy treatment detail: modality, total dose and duration

7. To enable a specific research study, we recommend collaborating with the principal investigator and statistician to identify the subset of variables needed to answer the study questions, and pivoting them into a square table, one patient per row, for the convenience of the statistician.

The model of data sharing described in this paper, depicted in Figure 6, is scalable to multiple institutions if all participating institutions adopt the same algorithm for generating the per-patient study identifier. Our development of the Oncoshare database represents one of several reported strategies for deriving data on cancer care variability and quality from the EMR. A relevant prior study of variability in breast surgery reported a chart abstraction approach [18], but did not extract as many treatment variables as we have done. We are now using Oncoshare to prepare a report on variability in breast cancer care and outcomes across multiple treatment modalities and

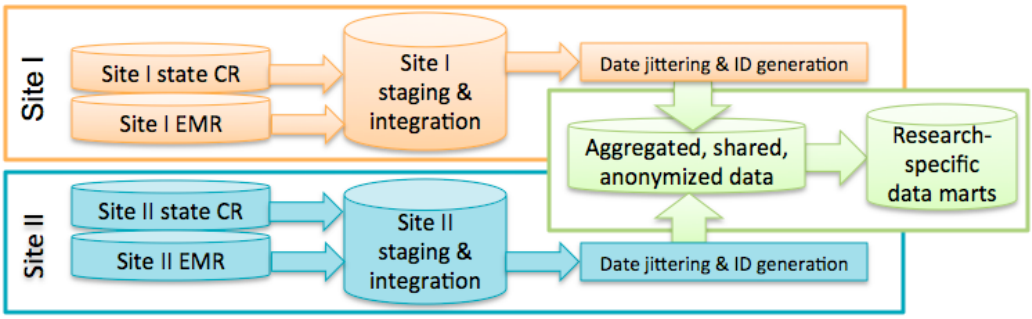

Figure 6: Recommended staging and data sharing architecture

institutions. In addition to the CER goals, the availability of a standardized database capturing breast cancer care has resulted in new informatics approaches to analyzing patterns of care, such as clustering by similarity of treatment history [19] and visualization and analysis of physician interactions across oncology specialties [20]. Ultimately, we have learned that integrating data from the EMR, the state cancer registry, and the SSA DMF provides a rich, validated resource that can answer many compelling research questions in cancer care.

Acknowledgments

This research was supported by funds provided by The Regents of the University of California, California Breast Cancer Research Program, Grant Number 16OB-0149. The opinions, findings, and conclusions herein are those of the authors and do not necessarily represent those The Regents of the University of California, or any of its programs. Funding for this research was also provided by the generous support of the Susan and Richard Levy Gift Fund. We thank the following people for discussions that contributed to this research: Scarlett Lin Gomez, Terri Owen, Peter Yu, Douglas Blayney, and Harold Luft.

References

1. Weinstein, MC, and Skinner, JA. Comparative Effectiveness and Health Care Spending — Implications for Reform. N Engl J Med 362:460-465, 2010.
2. Pace, WD, Cifuentes, M, Valuck, RJ, Staton, EW, Brandt, EC, and West, DR. An electronic practice-based network for observational comparative effectiveness research. Ann Intern Med 151(5):338-40, 2009
3. Tunis, SR, Benner, J and McClellan, M. Comparative effectiveness research: Policy context, methods development and research infrastructure. Statist. Med., 29: 1963–1976, 2010.
4. D'Avolio, LW, Farwell, WR, and Fiore, LD. Comparative effectiveness research and medical informatics. Am J Med. 123(12 Suppl 1):e32-7, 2010.
5. Weber SC, Lowe H, Das A, Ferris TA. A simple heuristic for blindfolded record linkage. J Am Med Inform Assoc doi:10.1136/amiajnl-2011-000329, 2012.
6. Komatsoulis GA, Warzel DB, Hartel FW, Shanbhag K, Chilukuri R, Fragoso G, Coronado S, Reeves DM, Hadfield JB, Ludet C, Covitz PA. caCORE version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability. J Biomed Inform. 2008 Feb;41(1):106-23. Epub 2007 Apr 2.
7. Cf. QuestGen systems blog post September 14 2011. http://www.quesgen.com/wp/three-ways-to-deal-with-hipaa-dates-in-de-identified-data-sets/
8. Neamatullah I, Douglass MM, Lehman LH, Reisner A, Villarroel M, Long W, Szolovits P, Moody GB, Mark RG, Clifford GD. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak. 2008; 8: 32.
9. Liu J, Erdal S, Silvey SA, Ding J, Riedel JD, Marsh CB, Kamal J.Toward a Fully De-identified Biomedical Information Warehouse. AMIA Annu Symp Proc. 2009; 2009: 370–374.
10. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR and Masys DR. Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. Clinical Pharmacology & Therapeutics (2008); 84, 3, 362–369 doi:10.1038/clpt.2008.89
11. Hernandez PN, Podchiyska T, Weber SC, Ferris TA, Lowe HJ. Automated Mapping of Pharmacy Orders from Two Electronic Health Record Systems to RxNorm within the STRIDE Clinical Data Warehouse. Proceedings of the 2009 AMIA Annual Symposium, 2009.
12. Jagsi R et al. Underascertainment of Radiotherapy in Surveillance, Epidemiology, and Results Registry Data, Published online June 29, 2011 in Wiley Online Library (wileyonlinelibrary.com)
13. As noted in Request for Confidential Data- California Cancer Registry http://www.ccrcal.org/pdf/Data_Statistics/CCRPoliciesProcedures_v05.1.pdf, linked from http;//www.ccrcal.org/Data_and_Statistics/Cancer_Data_for_Research.shtml
14. http://www.ntis.gov/products/ssa-dmf.aspx
15. May SG, Rendle K, Ventre N, Frosch DL, Kurian AW. A time to decide: patient perspectives on breast cancer treatment decision making. Oral presentation at the Qualitative Health Research Conference, Vancouver, BC, October 2011.
16. May SG, Rendle K, Halley M, Ventre N, Frosch DL, Kurian AW. Breast cancer survivors' post-treatment engagement with healthcare providers and recommendations for survivorship care. Submitted for presentation at the American Public Health Association Annual Meeting, San Francisco, CA, October 2012.

17. Paik S, Shak S, Tang G, et al: A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351:2817-26, 2004.
18. McCahill LE et al. Variability in Reexcision Following Breast Conservation Surgery, JAMA 307 (5): 467-475, 2012.
19. Bridewell, W, and Das, AK. Social network analysis of physician interactions: the effect of institutional boundaries on breast cancer care. Proceedings of the 2011 AMIA Annual Symposium, Washington, DC., 2011.
20. Lee, WN, Bridewell, W, and Das, AK. Alignment and clustering of breast cancer patients by longitudinal treatment history. Proceedings of the 2011 AMIA Annual Symposium, Washington, DC., 2011