Published in final edited form as:

Nurs Res. 2013 ; 62(6): 438–444. doi:10.1097/NNR.0000000000000002.

Linking Clinical Research Data to Population Databases

Linda S. Edelman, PhD, RN [Assistant Professor],

Robert S. and Beth M. Carter Endowed Chair in Nursing, University of Utah College of Nursing.

Jia-Wen Guo, PhD, RN [Assistant Professor],

Robert S. and Beth M. Carter Endowed Chair in Nursing, University of Utah College of Nursing.

Alison Fraser, MSPH [Senior Database Analyst], and

University of Utah Huntsman Cancer Institute.

Susan L. Beck, PhD, APRN, FAAN [Professor]

Robert S. and Beth M. Carter Endowed Chair in Nursing, University of Utah College of Nursing.

Abstract

Background—Most clinical nursing research is limited to funded study periods. Researchers can study relationships between study measures and long-term outcomes if clinical research data can be linked to population databases.

Objectives—The objective was to describe feasibility of linking research participant data to data from population databases in order to study long-term poststudy outcomes. As an exemplar, participants were linked from a completed oncology nursing research trial to outcomes data in two state population databases.

Methods—Participant data from a previously completed symptom management study were linked to the Utah Population Database and the Utah Emergency Department Database. The final dataset contained demographic, cancer diagnosis and treatment, and baseline data from the oncology study linked to poststudy long-term outcomes from the population databases.

Results—One hundred twenty-nine of 144 (89.6%) study participants were linked to their individual data in the population databases. Of those, 73% were linked to hospitalization records, 60% to emergency department visit records, and 28% were identified as having died.

Discussion—Study participant data were successfully linked to population databases data to describe poststudy emergency department visit and hospitalization numbers and mortality. The results suggest that data linkage success can be improved if researchers include linkage and human subjects protection plans related to linkage in the initial study design.

Keywords

cancer; outcomes; population database; randomized controlled trials; record linkage

The rapid advances in the early diagnosis and treatment of many diseases, including cancer, have resulted in decreased morbidity and mortality (Edwards et al., 2010). Increasingly, researchers are focusing on how health care interventions impact long-term clinical and epidemiological outcomes, such as well-being, survival, and utilization of resources (Hawkins & Robison, 2006). Unfortunately, the cost of following research participants over time is often prohibitive for all but the largest and best funded of studies. Therefore, the

majority of clinical research studies are limited in the length of follow-up time, as well as the long-term outcomes measured.

One possible way to follow research participants over long periods of time is to link their clinical research records to data in large population databases, clinical registries, and administrative databases. These large databases are often utilized for research; however, the type and quality of data varies greatly amongst them. Therefore, linkages among multiple databases are increasingly used to merge clinical data with administrative data to provide the power and depth of data needed to address clinical research questions extended through time (Dokholyan et al., 2009; Pasquali et al., 2013). Linking these large databases to clinical research data could be used to study outcomes such as health care utilization and survival over longer periods of time without having to track individual study participants themselves. Doing so requires that critical issues related to privacy and protection of human research subjects be addressed. Further, potential impediments to such linkages exist, including availability and accuracy of variables that could be used to correctly link an individual's record in one database to other databases, and the technology available to perform complex data linkages.

The purpose of this research was to test the feasibility of linking clinical research data to large population databases in order to explore the relationships of clinical research findings with long-term outcomes. As an exemplar, we used data from a previously conducted clinical research study of a nursing intervention to improve cancer-related fatigue which we linked to two different population databases maintained in Utah. The aim was to explore the relationship between baseline data at the time of cancer diagnosis and long-term outcomes, such as emergency department (ED) visits, hospitalizations, and survival. Here, the study (a) describes the data linkage process; (b) examines the feasibility of creating a linked dataset to explore poststudy outcomes; and (c) provides recommendations for researchers designing clinical research studies who are considering long-term plans for linking study data to population databases. (An extensive substantive report of findings from the linkage of our clinical research data with poststudy outcomes is beyond the scope of this paper.)

Methods

This feasibility study linked participants from an intervention study to improve cancer-related fatigue conducted a decade ago to the Utah Population Database and the state Emergency Department Database.

Databases

Study database—Data from the Energy Conservation and Activity Management (ECAM) for Patients with Cancer-related Fatigue trial were used as the exemplar. The ECAM trial was a multicenter, multistate study funded by the National Institutes of Health (R01 NU04573; Barsevick, PI) that tested a nursing intervention focused on helping patients to conserve energy as a strategy for managing fatigue during cancer treatment. The main findings of this study have been reported elsewhere (Barsevick et al., 2004; Barsevick, Dudley, & Beck, 2006). Participants enrolled in the study between September 1999 and October 2001 at the Huntsman Cancer Institute (HCI) in Salt Lake City, UT and Fox Chase Cancer Center in Philadelphia, PA. All patients were initiating chemotherapy, radiation therapy, or combination therapy. The dataset for this project of 151 participants enrolled at HCI included demographic (including age at enrollment) and clinical information, but did not contain names or date of birth. The original study also had consent forms with handwritten signatures and dates, and a separate file linking study ID to medical record numbers in the oncology clinic or radiation therapy department.

Utah Population Database (UPDB)—The UPDB is a research resource at the University of Utah containing Utah population data, including: vital statistics (e.g. birth, marriage and death certificates); cancer registry data from both Utah and Idaho; and driver's license data. It also includes claims data from statewide inpatient hospital discharge records, as well as ambulatory surgery records from hospital outpatient departments and ambulatory surgery centers ("Utah Population Database", *n.d.*). The UPDB contains a master subject index that allows for cross linkage with health care administrative records of all patient encounters maintained in electronic data warehouses (DuVall, Kerber, & Thomas, 2010).

Access to the UPDB is governed by the Utah Resource for Genetic and Epidemiologic Research (RGE), which contracts with UPDB data contributors, such as hospital systems and the Department of Health, and sets conditions for data use and reviews requests to access UPDB data. Personal information can be used by UPDB staff to match individuals from two or more datasets, and then removed from the final linked dataset provided to researchers (Wylie & Mineau, 2003).

For this study, the ECAM participant data were linked to the UPDB using the best identifying information available from all sources. Three databases contained within the UPDB primarily provided data for the study. The Utah Cancer Registry contains statewide information from individuals diagnosed with cancer since 1966. The vital records death certificates database contains cause of death information. The hospital discharge database contains statewide population-based information about hospitalizations in Utah.

Emergency Department Database (EDDB)—The EDDB is maintained by the Department of Health and contains information on all emergency department visits to Utah hospitals. Data of interest included visit information such as date, length of stay, discharge status, and treatment outcomes. The Intermountain Injury Control Research Center at the University of Utah has administrative access to the EDDB and, after approval from the Department of Health IRB, performed the linkage between the EDDB records and ECAM participant data.

Human Subjects Research Protection

Because linking ECAM participant data to other databases was not included in the original ECAM study design, there was concern about confidentiality and privacy issues. In addition, the UPDB and the EDDB are under the oversight of two different committees. Therefore, work was done in tandem with three oversight committees—the University of Utah Institutional Review Board (IRB), the Resource for Genetic and Epidemiologic Research (RGE) and the Utah Department of Health IRB—to ensure that research participant personal health information (PHI) was protected and that regulations for utilizing UPDB and EDDB records in research were followed.

The University of Utah IRB reviews all research projects that involve humans to ensure that they comply with University policy and local, state, and federal laws and regulations regarding research. University of Utah IRB approval requires RGE approval for use of UPDB data. The RGE governs access to UPDB data and research, and works in tandem with the University of Utah IRB to ensure that human subjects data are protected. Waivers of Consent and Authorization were requested from the University of Utah IRB, which detailed how data identifiers would be protected. A component of RGE approval is review of the research by the Health Data Committee, a Department of Health committee established under Utah Health Code (§26–33a) with oversight of the use and sharing of health care data collected by the Department of Health. Approval to access EDDB records was obtained from the Utah Department of Health IRB, which promotes ethical research and the protection of confidentiality and privacy of participants.

Once RGE and Utah Department of Health IRB approvals were obtained, the University of Utah IRB approved the study as a minimal risk study because study investigators never had access to identifying information other than what they possessed in the original ECAM study dataset. Personal health information contained within the UPDB and EDDB that was used to identify and link records was only available to database administrators. After the record linkages were complete, all identifying information was removed prior to returning the linked data to the study investigators. This manuscript was submitted to the directors of the Health Data Center (Department of Health), RGE, and the Utah Cancer Registry prior to submission for publication.

Data Linkage

The ECAM study dataset and consent forms were used to generate a data file for the University of Utah Data Warehouse, which maintains medical records of patients treated within the University of Utah Health Care system. Data Warehouse administrators matched medical record numbers provided by the ECAM dataset to two different patient databases that were in use at the time of the original study. Records with medical record number matches were confirmed using gender and estimated year of birth, and then forwarded to UPDB administrators for linkage to the UPDB datasets used for this study. Records without medical record number matches were manually linked to the UPDB using the other linkage variables.

Once the UPDB linkage was complete, the following information from the linked records was sent to the Intermountain Injury Control Research Center for linkage to EDDB records: names, birth dates, study recruitment dates, diagnosis dates, death dates, and zip codes. Probabilistic linkage using LinkSolv record linkage software (Strategic Matching, Inc., Morrisonville, N.Y.) was used to link UPDB records to all EDDB records for visits after the date of cancer diagnosis. Probabilistic linkage is a method that links computerized records from two or more disparate data sources by comparing common linkage variables shared by each to make a judgment that records refer to the same person or event within the different datasets (Edelman, Cook, & Saffle, 2009; Lee et al., 2006). Pairs of UPDB and ED records that had a 90% or greater probability of being a true match were linked. The final study dataset, which was a result of the linkage of the ECAM, UPDB, and EDDB databases, contained the information shown in Table 1.

Poststudy Outcomes

Mortality—An ECAM participant was considered deceased if there was a Utah vital records death record, Social Security death index record, or death date on a Utah Cancer Registry record for that individual between time of study enrollment and December 31, 2010.

Health care utilization—The number of emergency department visits and hospitalizations were determined by the number of individual visits a participant linked to in the EDDB, or in the hospital discharge database contained within the UPDB between the time of study enrollment and December 31, 2010.

Data Analysis

Data for this study were analyzed using the Statistical Package for the Social Sciences for Windows (Version 19.0; SPSS Inc., Chicago, IL). Descriptive statistics (means, standard deviation, and percentages) were used to describe demographics of the study sample and long-term outcomes (mortality and health care utilization, including emergency department visits and hospitalizations).

Results

Data Linkage

A schematic of the data linkage process is shown in Figure 1. There were ECAM data for 151 participants; however, seven participants either dropped out of the study or had incomplete data. Therefore, the results of the linkage of 144 participants to the UPDB and EDDB using an iterative process are reported. In the first step, 65 of 144 (45.1 %) ECAM study patients were matched to records in the Data Warehouse and assigned distribution numbers that link Data Warehouse records to the UPDB. In 10 potential cases, the Data Warehouse provided additional information that allowed matching these 10 study records with the UPDB. Now, 75 of 144 (52.1%), with names, were linked to UPDB records.

Then, a second strategy was tried. Very few ECAM records contained names (15/144); but the ECAM consent forms had participant signatures which were deciphered as best possible. Deciphered names were matched to names of individuals in the UPDB who had a cancer record with a diagnosis date in proximity to the study consent recruitment date and clerically reviewed. This liberal matching resulted in only 25 of 144 (17.4%) of ECAM participants who were not linked to either the ECAM data or the UPDB. In order to confirm positive matches, the UPDB administrator hand-matched ECAM consent form recruitment dates to the ECAM database, using other linkage variables, such as date of enrollment, gender, age, and cancer diagnosis to confirm a match. As a result, 129 of 144 ECAM participants were linked to the UPDB for a final linkage yield of 89.6%. All matches contained Utah Cancer Registry records, 94 (73%) linked to at least one hospital discharge record, and 31 (24%) linked to death certificate data.

The next step in the linkage process was to connect the 129 participants matched to UPDB records to the EDDB, which contains information on emergency department visits that do not result in hospitalization. Because the EDDB is not maintained within the UPDB, and involves different database administrators, identifying information (name, date of birth, Utah Cancer Registry dates of diagnosis date and ECAM recruitment dates) of ECAM participants was used to link EDDB data.

Sample Characteristics

Demographics of the 129 ECAM participants linked to the UPDB and or EDDB are shown in Table 2. Participants were predominantly White (94.6%) and female (87.6%). A vast majority were diagnosed with breast cancer (76.0%) and undergoing radiation (50.4%) or chemotherapy (45.0%).

Poststudy Outcomes

Sixty percent of participants had at least one emergency department visit ($M = 1.72$, $Mdn = 1.00$, range = 0–27); 49% of those who had ED visits had more than one visit (Table 3). The majority of participants (72.9%) were admitted to a Utah hospital at least once ($M = 2.16$; range 0–11 visits); 67% of participants who were admitted to the hospital had more than one hospitalization. As of December 31, 2010, 36/129 (27.9%) of ECAM participants were known to be deceased; 31 participants had death records and another five participants were documented as deceased in the Utah Cancer Registry datasets contained within the UPDB.

Discussion

Clinical investigations in nursing are often time-limited and focused on a specific phase of diagnosis or treatment. Longitudinal studies are expensive, and study attrition over time can

compromise data integrity. Thus, studies that extend beyond a brief clinical trial phase usually require large numbers of participants and assurance of continued funding.

Population databases can include local, state, and national health care utilization and vital records databases, as well as hospital and administrative databases and disease registries that contain health outcomes data. These databases can provide researchers with clinical research study participant information on a range of health care utilization and long-term outcomes at different time points after a study has ended, saving research time and costs, and reducing participant burden (Moriarty et al., 1999). As illustrated, successful linking of clinical research data to population databases depends availability and quality of both the clinical research and population database data, the participant identifiers available, and the variables shared between databases (Moriarty et al., 1999). All these factors can affect the validity of findings based on the linked data.

Feasibility of linking clinical research data to population databases was enhanced by an established record of collaborations between health sciences researchers at our University and the Utah Department of Health, which maintains a number of state population databases (Edelman et al., 2009). The University of Utah is the home to the UPDB, one of the world's richest sources of in-depth information on more than 7.2 million individuals who can be used to support research on genetics, epidemiology, demography, and public health (Wylie & Mineau, 2003). The UPDB holds records from a number of contributors, including the Utah Department of Health (births, deaths, marriages/divorces, hospitalizations, and ambulatory surgery), family history records, driver's licenses, and voter registration records ("Data," 2013). Researchers have utilized UPDB data retrospectively to identify individuals with different cancer phenotypes and genotypes (Albright, Teerlink, Werner, & Cannon-Albright, 2012; Chow et al., 2009; Lin et al., 2011; Mueller et al., 2009). The UPDB has also been used to link research subjects to relatives with data in the UPDB to study the effects of BRCA1/2 mutations on female fertility (Smith, Hanson, Mineau, & Buys, 2012). UPDB data has recently been linked to data from a completed population-based autism prevalence study in order to examine mortality and causes of death (Bilder et al., 2012); unlike the feasibility study we reported, the primary research study data included consistent identifying information.

The results of this study demonstrate the feasibility of linking completed oncology research participant data to large population databases to answer questions about long-term outcomes. Almost 90% of site data were linked from a multicenter oncology study (ECAM) completed a decade ago to the UPDB and the EDDB. However, the lack of consistent identifying information about the research participants required manual record matching, which is difficult and time-consuming. Had the dataset of ECAM participants been large, this method would not have been feasible.

The site for the feasibility study is a regional cancer center and treats patients from a multistate area. Of the 129 ECAM participants linked to the UPDB and EDDB, 13 (10%) were residents of a neighboring state. Because the UPDB and Utah Emergency Department Databases only include information on individuals who were treated at Utah hospitals, there was no information on emergency department visits and hospitalizations of participants when they occurred outside of Utah, which may have resulted in an under reporting of ED visits and hospitalizations in this study. Conversely, the Utah Cancer Registry does collect death certificate data from other states, so data about mortality events were not impacted by whether the death occurred in Utah.

By linking data from the ECAM oncology study to databases within the UPDB and the EDDB, it was possible to describe the number of emergency department visits,

hospitalizations, and deaths experienced by study participants during the time following completion of the study. Poststudy, long-term follow up of research participants can provide important information on survivorship, quality of life, general health, and health care utilization. Unless a study has a plan to engage participants, long-term, follow up can be difficult and expensive. Linking participants to available population databases at different points in time can be an affordable way to acquire long-term outcome data.

This exemplar demonstrated that linking clinical research data to population databases in order to study long-term outcomes, poststudy closure is feasible. Large databases, such as administrative, state and other patient or disease-oriented databases, may contain a plethora of information that could be used to study research participants longitudinally (Magee, Lee, Giuliano, & Munro, 2006). Moriarty et al. (1999), identified a number of advantages to using large national databases for research of families which can be generalized to most studies. These include the potential savings in time, costs and resources, decreased subject burden, and the inclusion of more variables (Moriarty et al., 1999). As discussed above, database linkages can be compromised by data availability and quality, which can contribute to bias and may impact the validity of the findings (Moriarty et al., 1999).

Recommendations

Recruitment to clinical studies is costly. Poststudy follow up of participants should be considered and facilitated to leverage the value of project data. It is recommended that researchers incorporate anticipated data linkage plans when designing clinical studies rather than doing so retrospectively when the study has been completed. In doing so, researchers should take the following into consideration:

Always consider human subjects research and ethical issues—Be familiar with and incorporate your institutional review board's guidelines and requirements of other oversight committees associated with the databases into your study design. For example, if you know which databases you may want to link study participant data to in the future, permission to do so should be included in the study consent form. Recognize that working with multiple oversight committees takes considerable time and effort. Involve the committees early in the research design process. In this exemplar, approval was received from three different oversight committees. The process went smoothly because key stakeholders were included early in the conceptualization and design of the data linkage process.

Retain allowable research participant identifying information in a secure fashion—This study demonstrated the need to keep identifying information in a separate file, but linked to the clinical research dataset with a unique study ID. Some institutions might require that this data be securely retained by an institutional research administration group or a central repository for multicenter studies. If maintaining personal identifying information is not possible, consider the use of indirect patient identifiers such as birthdates, gender, and admission and discharge dates that could be used to link research data to other databases when allowable. Indirect identifiers have been used to successfully link clinical registry data with administrative databases (Pasquali et al., 2010).

Know what personal identifying information is sufficient to provide the best linkage yield—Consider what identifying numbers are most likely to be retained over time in other databases. When allowable, full name, gender, and birth date are key linking variables. Consent forms should have the signature and the name printed underneath. Additional information such as residential address, medical record numbers, and Social

Security numbers can be helpful in confirming potential data links when participants provide them and their use is approved as part of the linkage process.

Identify potential long-term outcomes in a study—When designing a study, identify potential long-term outcomes that may be of interest in the future, what databases might house those outcome variables, and if the variables are congruent with the study's conceptual framework (Magee et al., 2006). Meeting with database administrators at this stage, to discuss the administrative guidelines for accessing the data, will ensure ease of linkage in the future.

Conclusions

Studying long-term clinical research outcomes can be complicated by the difficulty and expense of following research participants over time. In this paper, the feasibility of using probabilistic linkage to join a small clinical research database with large population databases to explore the relationships of clinical research findings with poststudy outcomes was demonstrated. Because the linkage to population databases was conducted many years after the clinical study was completed, and identifying information had been purged or was inadequate, the resulting linkage process was difficult. Therefore, researchers should consider a number of issues when considering the use of population databases to study long-term health outcomes. Most importantly, including poststudy linkage plans in the initial study design will simplify the linkage process and improve the linkage yield. Including a linkage plan in the original study design creates the opportunity to obtain explicit consent from participants to use specific data items for linkage with population databases, which can support consideration of the linkage project by committees charged with oversight and approval. In addition, it is important to establish relationships with database administrators, understand the responsibilities of database administrators, and archive clinical research data. Protection of participant confidentiality is paramount at every step of the process. With planning, population databases can be a rich and cost-effective source of data on the long-term outcomes of participants in clinical research studies.

Acknowledgments

The authors acknowledge a University of Utah College of Nursing Innovative Seed Grant made possible by a gift from Mr. and Mrs. Richard McGillis; the Utah Cancer Registry, which is funded by Contract No. HHSN261201000026C from the National Cancer Institute's SEER Program with additional support from the Utah State Department of Health and the University of Utah; partial support for all datasets within the Utah Population Database is provided by the Huntsman Cancer Institute/Huntsman Cancer Foundation and by the National Cancer Institute Cancer Center Support Grant P30CA042014; and the original research was funded by the National Institute for Nursing Research R01 NU04573 (Andrea Barsevick, PI).

References

- Albright F, Teerlink C, Werner TL, Cannon-Albright LA. Significant evidence for a heritable contribution to cancer predisposition: A review of cancer familiarity by site. *BMC Cancer*. 2012; 12:138. [PubMed: 22471249]
- Barsevick AM, Dudley W, Beck S, Sweeney C, Whitmer K, Nail L. A randomized clinical trial of energy conservation for patients with cancer-related fatigue. *Cancer*. 2004; 100:1302–1310. [PubMed: 15022300]
- Barsevick AM, Dudley WN, Beck SL. Cancer-related fatigue, depressive symptoms, and functional status: A mediation model. *Nursing Research*. 2006; 55:366–372. [PubMed: 16980837]
- Bilder D, Botts EL, Smith KR, Pimentel R, Farley M, Viskochil J, Coon H. Excess mortality and causes of death in autism spectrum disorders: A follow up of the 1980s Utah/UCLA autism epidemiologic study. *Journal of Autism and Developmental Disorders*. 2013; 43:1196–1204. [PubMed: 23008058]

- Chow EJ, Kamineni A, Daling JR, Fraser A, Wiggins CL, Mineau GP, Mueller BA. Reproductive outcomes in male childhood cancer survivors: A linked cancer-birth registry analysis. *Archives of Pediatrics and Adolescent Medicine*. 2009; 163:887–894. [PubMed: 19805706]
- Dokholyan RS, Muhlbaier LH, Falletta JM, Jacobs JP, Shahian D, Haan CK, Peterson ED. Regulatory and ethical considerations for linking clinical and administrative databases. *American Heart Journal*. 2009; 157:971–982. [PubMed: 19464406]
- DuVall SL, Kerber RA, Thomas A. Extending the Fellegi-Sunter probabilistic record linkage method for approximate field comparators. *Journal of Biomedical Informatics*. 2010; 43:24–30. [PubMed: 19683070]
- Edelman LS, Cook L, Saffle JR. Using probabilistic linkage of multiple databases to describe burn injuries in Utah. *Journal of Burn Care & Research*. 2009; 30:983–992. [PubMed: 19826268]
- Edwards BK, Ward E, Kohler BA, Ehemann C, Zaubler AG, Anderson RN, Ries LA. Annual report to the nation on the status of cancer, 1975–2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates. *Cancer*. 2010; 116:544–573. [PubMed: 19998273]
- Hawkins MM, Robison LL. Importance of clinical and epidemiological research in defining the long-term clinical care of pediatric cancer survivors. *Pediatric Blood & Cancer*. 2006; 46:174–178. [PubMed: 16333860]
- Lee MK, Lim SW, Yang H, Sung SH, Lee HS, Park MJ, Kim YC. Osteoblast differentiation stimulating activity of biflavonoids from *Cephalotaxus koreana*. *Bioorganic & Medicinal Chemistry Letters*. 2006; 16:2850–2854. [PubMed: 16574412]
- Lin W-Y, Camp NJ, Cannon-Albright LA, Allen-Brady K, Balasubramanian S, Reed MWR, Cox A. A role for XRCC2 gene polymorphisms in breast cancer risk and survival. *Journal of Medical Genetics*. 2011; 48:477–484. [PubMed: 21632523]
- Magee T, Lee SM, Giuliano KK, Munro B. Generating new knowledge from existing data: The use of large data sets for nursing research. *Nursing Research*. 2006; 55(2 Suppl):S50–S56. [PubMed: 16601635]
- Moriarty HJ, Deatrick JA, Mahon MM, Feetham SL, Carroll RM, Shepard MP, Orsi AJ. Issues to consider when choosing and using large national databases for research of families. *Western Journal of Nursing Research*. 1999; 21:143–153. [PubMed: 11512173]
- Mueller BA, Chow EJ, Kamineni A, Daling JR, Fraser A, Wiggins CL, Drews-Botsch C. Pregnancy outcomes in female childhood and adolescent cancer survivors: A linked cancer-birth registry analysis. *Archives of Pediatrics and Adolescent Medicine*. 2009; 163:879–886. [PubMed: 19805705]
- Pasquali SK, Jacobs JP, Shook GJ, O'Brien SM, Hall M, Jacobs ML, Li JS. Linking clinical registry data with administrative data using indirect identifiers: Implementation and validation in the congenital heart surgery population. *American Heart Journal*. 2010; 160:1099–1104. [PubMed: 21146664]
- Pasquali SK, Peterson ED, Jacobs JP, He X, Li JS, Jacobs ML, Mayer JE. Differential case ascertainment in clinical registry versus administrative data and impact on outcomes assessment for pediatric cardiac operations. *Annals of Thoracic Surgery*. 2013; 95:197–203. [PubMed: 23141907]
- Smith KR, Hanson HA, Mineau GP, Buys SS. Effects of BRCA1 and BRCA2 mutations on female fertility. *Proceedings of the Royal Society: Biological Sciences*. 2012; 279:1389–1395. [PubMed: 21993507]
- “Utah Population Database”. (*n.d.*). Retrieved from <http://www.huntsmanccancer.org/research/shared-resources/utah-population-database/data>.
- Weinert C. A social support measure: PRQ85. *Nursing Research*. 1987; 36:273–277. [PubMed: 3306610]
- Wylie JE, Mineau GP. Biomedical databases: Protecting privacy and promoting research. *Trends in Biotechnology*. 2003; 21:113–116. [PubMed: 12628367]

Figure 1.
Process of linking records from clinical research study to the final study outcomes database.

Table 1

Variable Sources and Linkages

Linkage			Potential Baseline Data		Poststudy Outcomes	
Variables	Databases		Variables	Databases	Variables	Databases
Medical record number	ECAM → UPDB		Perceived social support ^a	ECAM	Hospitalizations ^b	UPDB
Name	ECAM ↔ UPDB ↔ EDDB		Age	ECAM, UPDB, EDDB	ED visits ^c	EDDB
Age-specific time points	ECAM ↔ UPDB ↔ EDDB		Gender	ECAM, UPDB, EDDB	Mortality ^d	UPDB
Gender	ECAM ↔ UPDB		Marital status	ECAM		
Date of birth	UPDB ↔ EDDB		Comorbidities	ECAM		
Death dates	UPDB ↔ EDDB		Cancer type and stage	ECAM		
			Treatment type	ECAM		
			Symptom severity	ECAM		

Note. ECAM = Energy conservation for patients with cancer-related fatigue trial. EDDB = Emergency Department Database. UPDB = Utah Population Database. ED visit = Emergency Department.

^aPersonal Resource Questionnaire (PRQ85)(Weinert, 1987)

^bDate, ICD-9 code, length of stay, disposition, hospital location, number of hospitalizations

^cDate, ICD-9 code, disposition, number of ED visits

^dMortality status, death (date, location, cause)

Table 2

Participant Demographics and Post-study Outcomes

Characteristic	<i>n</i>	(%)
Gender		
Female	113	(87.6)
Male	16	(12.4)
Ethnicity		
Non-Hispanic	125	(96.9)
Hispanic	4	(3.1)
Race		
White	122	(94.6)
Black	1	(.8)
Asian or Pacific Islander	2	(1.6)
Other	4	(3.1)
Diagnosis		
Breast	98	(76.0)
Lymphoma	12	(9.3)
Lung	10	(7.8)
Cervical	4	(3.1)
Colorectal	4	(3.1)
Testicular	1	(.8)
Stage level		
Stage 0	6	(4.9)
Stage I	36	(29.3)
Stage II	49	(39.8)
Stage III	25	(20.3)
Stage IV	6	(4.9)
Treatment		
Radiotherapy	65	(50.4)
Chemotherapy	58	(45.0)
Combination therapy	6	(4.7)
Post-study death	36	(27.9)
	<i>M (SD)</i>	<i>Range</i>
Age (year)	54.48 (12.68)	18–83
Total comorbidities ^a	.88 (.93)	0–3
Number of hospitalizations	2.16 (2.33)	0–11
Number of ED visit	1.72 (3.62)	0–27

Note. *N* = 129.

^aPossible range: 0 – 8.