SF 424 R&R and PHS-398
SpecificTable of Contents

PI: **Kantor, Elizabeth David**	Title: Obesity, chemotherapy dosing, and breast cancer outcomes	
	FOA: PA16-160	
	FOA Title: NIH Research Project Grant (Parent R01)	
	Organization: SLOAN-KETTERING INST CAN RESEARCH	
	Department: Epidemiology and Biostatistics	
Senior/Key Personnel:	*Organization:*	*Role Category:*
Elizabeth Kantor PhD	Sloan Kettering Institute for Cancer Research	PD/PI
Lawrence Kushi ScD	Kaiser Foundation Hospitals	Co-Investigator
Erin Bowles MPH	Group Health Cooperative	Co-Investigator
Denise Boudreau PhD	Group Health Cooperative	Co-Investigator
Elisa Bandera MD	Rutgers, The State University of New Jersey	Co-Investigator
Candyce Kroenke ScD	Kaiser Foundation Hospitals	Co-Investigator
Tatjana Kolevska MD	Kaiser Foundation Hospitals	Co-Investigator
Jennifer Griggs MD	University of Michigan	Consultant

OMB Number: 4040-0001
Expiration Date: 10/31/2019

RESEARCH & RELATED OTHER PROJECT INFORMATION

1. Are Human Subjects Involved?* ● **Yes** ❍ **No**

1.a. If YES to Human Subjects

Is the Project Exempt from Federal regulations? ❍ Yes ● No

If YES, check appropriate exemption number: 1 __ 2 __ 3 __ 4 __ 5 __ 6If

NO, is the IRB review Pending? ● **Yes** ❍ **No**

IRB Approval Date:

Human Subject Assurance Number FW00004998

2. Are Vertebrate Animals Used?* ❍ **Yes** ● **No**

2.a. If YES to Vertebrate Animals

Is the IACUC review Pending? ❍ Yes ❍ No

IACUC Approval Date:

Animal Welfare Assurance Number

3. Is proprietary/privileged information included in the application?* ❍ **Yes** ● **No**

4.a. Does this project have an actual or potential impact - positive or negative - on the environment?* ❍ **Yes** ● **No**

4.b. If yes, please explain:

4.c. If this project has an actual or potential impact on the environment, has an exemption been authorized or an environmental assessment (EA) or environmental impact statement (EIS) been performed? ❍ Yes ❍ No

4.d. If yes, please explain:

5. Is the research performance site designated, or eligible to be designated, as a historic place?* ❍ **Yes** ● **No**

5.a. If yes, please explain:

6. Does this project involve activities outside the United States or partnership with international collaborators?* ❍ **Yes** ● **No**

6.a. If yes, identify countries:

6.b. Optional Explanation:

ABSTRACT

Epidemiologic studies have linked obesity to poor breast cancer outcomes, and it has been suggested that obese women may experience poorer outcomes, in part, due to inadequate dosing of cytotoxic agents among obese women. Specifically, most cytotoxic agents are dosed according to body surface area, and therefore, the larger the woman, the higher the absolute dose. However, evidence shows that clinicians are more likely to depart from recommended dosing among heavier women for fear of inducing chemotherapy-associated toxicity. In 2012, the American Society of Clinical Oncology (ASCO) released guidelines stating that obese women should be dosed according to their full body surface area, largely based on evidence that suggested that fully-dosed obese women do not appear to experience more toxicity than fully-dosed normal-weight women. However, the guidelines acknowledge that data are extremely limited with regard to more severe obesity and in the real-world context of comorbidities. Furthermore, these guidelines cite that this practice of 'dose reducing' obese women may be one reason contributing to the poorer outcomes observed in this group. However, to date, no empirical investigations have sought to determine if, and to what extent, dose-reduced chemotherapy may explain differences in breast cancer survival. These guidelines were met with some criticism, citing the need for further evidence, and data suggest continuing uncertainty about proper dosing of obese cancer patients. Understanding the drivers of dose reductions may help better inform our understanding of this practice and efforts to disseminate guidelines; however, we know little about factors driving dose intensity, and how these factors may vary by body size.

We therefore propose to address these gaps using data on nearly 34,000 Stage I-IIIA breast cancer patients diagnosed and treated at Kaiser Permanente Northern California and at Group Health. Specifically, we will use the rich data from these integrated healthcare delivery systems to examine the relationship between body size and dose intensity, and will further examine how the factors contributing to dose reductions vary by body size (Aim 1). We will also evaluate if, and to what extent, dose reductions mediate the association between obesity and breast cancer recurrence and survival (Aim 2). Lastly, we will evaluate the association between body size and toxicity among women identified as receiving the full BSA-determined dose of chemotherapy (Aim 3). Our findings will provide critical and timely information to support or to warrant modification of current recommendations for chemotherapy dosing for obese breast cancer patients. Given the high and increasing prevalence of obesity in the United States, it is critical that we improve our understanding of chemotherapy dosing. The knowledge gained from this study can be used to better inform optimal treatment for the estimated 102,000 obese women diagnosed with breast cancer each year in the United States.

PROJECT NARRATIVE

Epidemiologic studies have linked obesity to poor breast cancer outcomes, and it has been suggested that obese women may experience poorer outcomes, in part, due to dose reductions of chemotherapy drugs; this is because most cytotoxic agents are dosed according to body size, and clinicians may scale back the high doses administered to obese women due to concern about inducing toxicity. Given persisting uncertainty about dosing obese women, in a study population of nearly 34,000 women with breast cancer who were diagnosed and treated in two integrated healthcare delivery systems, Kaiser Permanente Northern California and Group Health, we propose to: i) examine the relationship between body size and dose intensity, and will further examine how the factors contributing to dose reductions vary by body size, ii) evaluate whether chemotherapy dose reductions mediate the associations between obesity and adverse breast cancer outcomes, and iii) evaluate the association between body size and toxicity among women identified as receiving the full body-size determined dose of chemotherapy. Addressing these questions will provide the evidence needed to better inform clinicians treating the 102,000 obese women diagnosed with breast cancer each year in the United States.

FACILITIES AND OTHER RESOURCES

Memorial Sloan Kettering Cancer Center

Memorial Sloan Kettering Cancer Center (MSK) is one of the world's premier cancer centers. The institution is a comprehensive cancer center whose purposes are the treatment and control of cancer, the advancement of biomedical knowledge through laboratory and clinical research, and the training of scientists, physicians and other health care workers. It consists of three corporations: Memorial Hospital for Cancer and Allied Diseases (MH), founded in 1884, is the clinical unit; Sloan Kettering Institute for Cancer Research (SKI), founded in 1945, is the basic research unit; and Memorial Sloan Kettering Cancer Center, founded in 1960, is the unit which formulates policies and long-range plans and coordinates the activities of the Hospital and the Institute. Memorial Hospital consists of 470 inpatient beds, and in 2014 there were more than 590,000 outpatient visits. More than 770 research protocols are active throughout the institution.

Dr. Craig B. Thompson serves as President and Chief Executive Officer of Memorial Sloan Kettering Cancer Center. Dr. José Baselga serves as Physician-in-Chief of Memorial Hospital. Dr. Joan Massagué serves as Director of the Sloan Kettering Institute. MSK is a National Cancer Institute-designated Comprehensive Cancer Centers, affirming our leadership in research, patient care and education. Located in the "research corridor" of Manhattan's Upper East Side, Sloan Kettering Institute enjoys a close collaboration with neighbors Weill Cornell Medical College at Cornell University and The Rockefeller University. This Tri-Institutional community is home to nine Nobel Prize winners.

Department of Epidemiology and Biostatistics

The Department of Epidemiology and Biostatistics at Memorial Sloan Kettering Cancer Center consists of a group of professional staff investigators, and a large team of support staff, under the leadership of Dr. Colin Begg. Departmental staff members include 35 MD and/or PhD level research faculty, 20 Master-level statisticians and epidemiologists, and 10 computing support staff members. Our investigators pursue research designed to better understand the causes of cancer, the quality of care delivery on a population basis, and research methodology. They also apply their specialized disciplinary skills to enhance the reach of multidisciplinary research projects on cancer diagnosis, treatment and prevention.

The Department runs several seminar series that have hosted many distinguished speakers. The Biostatistics Service Seminar series, held weekly, hosts both department faculty presentations as well as invited speakers from other institutions. The Health Outcomes Research Group hosts twice-monthly Research-in-Progress seminars led by members of the group or by invited external speakers. The Epidemiology Service meets biweekly to discuss research in progress, planned grant applications, and to convene a Journal Club; the Molecular Epidemiology laboratory holds weekly meetings as well.

Computer Environment

The Department operates two server clusters and approximately 180 workstations which utilize Windows, Apple, and Linux operating systems.

The Database Server Cluster supports the institutional Clinical Research Database (CRDB) with two HP servers. This cluster consists of two HP Integrity rx6600 servers, both with 3 Core 1.6GHz CPUs. Each has 32GB memory. This cluster is connected to a shared EMC V-MAX storage array containing 5.2TB of disk space in either mirrored or RAID5 configuration. It runs the HP-UX 11.31 operating system. The cluster is front-ended by three HP Itanium rx2660 application servers, each with 8GB memory, 2 Core 1.66GHz CPUs running HP-UX 11.31 operating system. The cluster is also front-ended by three IBM X-3650 servers. Each server has 4 Core Xeon X5450 CPUs and 16GB memory running Red Hat Enterprise Linux 5.3 operating system. Oracle 10g RAC database software and related tools are installed on the Database Server Cluster to support the CRDB. Access and security are restricted to the CRDB Database by use of Oracle's internal security. Access to CRDB severs is restricted to DBA and System Administrator.

The Computation Server Cluster supports the statistical analysis and research performed by departmental statisticians. The cluster consists of 15 servers: 10 HP Proliant DL360p Gen8 nodes with 2 six-core Xeon E5-

2640 2.5 GHz CPU and 64 GB of RAM, one DL360 G7 node with 2 six-core Xeon E2645 2.4 GHz CPU and 48 GB of RAM, three HP Proliant DL380 G5 nodes with 2 Intel quad-core Xeon 2.33GHz CPU and 32 GB RAM and an HP Proliant G7 MicroServer with 6 GB RAM that acts as the dedicated LDAP authentication node. These servers run Xubuntu (Linux) v12.04 (Precise Pangolin) operating system and support all open source statistical and programming software (R v3.0.1, C, Fortran etc.). SAS 9.2, Stata 12 and MATLAB 7.8 statistical programs are available on one node each. Sun Grid Engine has been installed in these servers which facilitates efficient and simultaneous processing of multiple jobs across the cluster. A HP X1600 G2 Network Storage Server with 24TB disk space serves the cluster to provide local storage needed for the analysis of large data from high throughput assays.

Approximately 180 personal computers are available to the department's staff. The minimum processor standard is 3.0GHz Pentium 4, with a minimum of 4 GB of memory. A mixture of Microsoft Windows XP Professional and Linux operating systems are used on department PCs. All PCs are members of the institutional domain and are connected to MSK's enterprise file and print servers. Access to the domain is restricted via hardware, user and group levels. This includes restricted access to data files, shared and non-shared, and restricted access to all servers and other PC's. Several popular statistical, database, spreadsheet and presentation programs such as Microsoft Office, SAS, R, Stata, Cytel, and S-Plus are installed on each PC's local hard disk, as is the client software for the CRDB and Oracle user and development tools. One scanner workstation uses Verity Teleform version 9 to assist with electronic conversion of data to the CRDB.

Daily backups are taken of all data on the UNIX systems and on the enterprise file servers. All users' personal and departmental shared network drives receive a backup nightly, with a seven day on-hand user available restore. Every workstation and windows server has virus protection and auto patch procedures in place to keep all systems up to date with patches and virus definitions. All computers in the department are connected, along with approximately 35 printers/copiers, to a local area network through Cisco 4506-E switches, using the 10/100/1000 (PoE) protocol.The switches are connected to the institutional fiber optic Ethernet backbone and thus to the Internet. A comprehensive firewall, spam filter and virus detection and prevention system has been installed on the MSK network to prevent unwanted intrusion of any Server or workstation. All MSK staff/users receive mandatory institutional trainings on Information Security and must adhere to MSK security policies and standards at all times.

Core facilities
MSK has full-time computer consultants, programmers, and data managers to assist all project personnel. The Bioinformatics Core has established a robust computer infrastructure at MSK, which includes database, web, email and High Performance Compute (HPC) servers. The Geoffrey Beene Translational Oncology Core Facility and the Genomics Core Lab --both located within Main Campus-- provide, among many other services, microarray and high volume sequencing support for research projects as needed. The Radiation Core Facility, also located across the street from the Molecular Epidemiology Laboratory in the Zuckerman Building, provides access to a Mark I Gamma irradiator for irradiation of control cells as needed. The Sterilization and Glassware facility provides sterilization and glass/plastic ware washing services to the Molecular Epidemiology Laboratory.

Libraries
MSK investigators have access to library collections at three neighboring institutions: the MSK Library; the Weill Cornell University Medical School Library; and the Rockefeller University Library. The Department of Epidemiology and Biostatistics maintains its own library of specialized journals going back many years at the departmental offices on Lexington Avenue.

KAISER PERMANENTE NORTHERN CALIFORNIA – FACILITIES AND OTHER RESOURCES

THE KAISER PERMANENTE NORTHERN CALIFORNIA SETTING

Kaiser Permanente Northern California (KPNC), established in 1945, is one of the largest and oldest integrated health care delivery organizations in the U.S. One of seven regional entities nationwide, KPNC has a service area that encompasses the San Francisco Bay Area and the Central Valley from the Sacramento area in the north to Fresno in the south. KPNC provides care to approximately 3.8 million members – people who receive care through its health plan – or approximately 31.4 percent of the population in its service area. A comprehensive health care provider, KPNC owns and operates 21 hospitals and over 200 outpatient clinics

As an integrated health care system, KPNC brings together three separate organizations: the Kaiser Foundation Health Plan (KFHP), which is the insurance provider; Kaiser Foundation Hospitals (KFH), which owns and operates KP facilities; and The Permanente Medical Group (TPMG), the physician-owned practice group that provides care at KFH facilities to members who have health care coverage through KFHP.

THE DIVISION OF RESEARCH

The Division of Research (DOR) was established in 1961 as a department within The Permanente Medical Group of Kaiser Permanente Northern California. Its missions are to conduct public-domain epidemiological and health services research, and to serve as a principal consulting resource for health services and clinical studies for Kaiser Permanente Northern California (KPNC). Unlike internal analytic groups of KPNC, the DOR's research agenda is shaped primarily by the interests and initiative of its investigators. However, as part of KPNC, DOR researchers have direct access for research purposes to the administrative and clinical databases and electronic medical records of the organization; these are described in further detail below. In addition, as part of a large health care system, DOR research projects often include clinical partners and can have direct translational impact into clinical practice.

With its focus on conducting public-domain, peer-reviewed research, with dissemination of research results in peer-reviewed publications and presentations at professional conferences, the DOR's academic orientation and its organization is similar to those of large departments or moderate-sized schools of public health in academic health centers. Its 60 investigators and over 450 staff members are supported principally through externally-funded grants and contracts with an annual operating budget exceeding $80 million, the vast majority of which comes from externally-funded grants and contracts. The principal funding sources include Federal agencies, especially the National Institutes of Health but also including the Agency for Health Care Research and Quality, the Centers for Disease Control and Prevention, and other agencies. Other funding sources include the State of California and its Tobacco-related Disease Research and Breast Cancer Research Programs; non-profit organizations such as the American Diabetes Association or the American Heart Association; private foundations; industry funding, such as pharmaceutical or biotechnology companies; and directed funding from KP organizational sources to carry out specific research or infrastructure projects. Only a small proportion (about 5 percent) of the total operating budget of the DOR comes as flexible funding from the parent organization.

DOR investigators conduct research in a wide range of areas across the disciplines of epidemiology and health services research, from autism to Alzheimer's disease. Examples of research areas in which several investigators share complementary interests include cancer, cardiovascular disease, diabetes and metabolic disorders, maternal and child health, drug and alcohol research, health disparities, comparative effectiveness, and delivery science. Researchers and staff encompass a wide range of expertise, including survey methodology, biostatistics, psychology and behavioral sciences, medical anthropology, data management, research grants and contracts administration, and numerous clinical specialties.

Collaborative research ties through specific research projects exist with local institutions such as the Universities of California at Berkeley, San Francisco, and Davis, Stanford University, the California State Department of Public Health, as well as with many other institutions across the U.S.

Office

The DOR occupies a five-story office building in downtown Oakland, CA, and a floor of clinical and laboratory space in a thirteen-story building on the Kaiser Permanente Medical Center Oakland campus. The DOR also has a biorepository located in Berkeley, CA, on part of the regional laboratory facilities campus for KPNC. There is ample office space and resources (photocopying machines, computer workstations connected to a local area network, high-speed internet access, etc.) for all staff. Additional resources include a data entry

department, medical library, a medical record coder unit for chart review and abstraction, programmers and data analysts, information technology support staff, and experienced telephone interviewers.

Additional facilities available to DOR research projects include a full complement of health care services based both at Kaiser Permanente's regional level and at each individual medical center. These include comprehensive Health Education departments and Health Education libraries, an extensive regionalized data management system with consultants available to users, and Public Relations and Marketing representatives whose responsibilities are to maintain professional contact with major employers in the community, municipal government, and a large number of community organizations, with the goal of increasing public awareness of the various services and activities within Kaiser Permanente of Northern California.

Computer and Data Management Resources

The DOR maintains and controls a technology infrastructure that is comprised of, but not limited to, servers, data storage, and networking equipment housed in a data center. DOR also maintains and controls approximately 800 devices comprised of desktops, laptops, printers, mobile devices, and miscellaneous other devices. The DOR IT department employs various individuals with skillsets to implement, maintain, and address hardware and software requests or issues.

The technology infrastructure sits on various platforms. These platforms include Microsoft Windows, UNIX, and Linux. The DOR uses HP and Oracle servers for its primary systems. These systems have a minimum of 32 to 64 GB of RAM, and as high as 128 GB of RAM, and have built-in redundancy such as dual network cards and dual power supplies. A minimum of 16 cores are purchased for the systems. The DOR also uses virtualization to provide a scalable environment. Virtualization provides research investigators and staff with an environment that is fully redundant, eliminating unwanted outages. In this way, the DOR is responsible for approximately 120 physical servers and over 300 virtual servers.

Data for the systems are stored on SAN or NAS devices. Approximately 400 TB of usable data storage is available to DOR investigators. Standard storage vendors such as EMC, Oracle, or Network Appliance are used. Data drives for these storage systems are based on performance requirements. For example, SSD drives are used for high input/output utilization. The combination of these systems allows investigators to obtain the highest quality pieces of technology as required. Data storage systems are connected to servers via fiber optics or high-speed networking.

Servers, data storage systems, and backup systems connect via Cisco network switches at 10 gigabit speeds or faster. Desktops and laptops connect to the system via 1 gigabit connections. The network infrastructure is protected by a redundant firewall. The firewall also provides a DMZ for housing Internet-facing applications. Additional security is provided by an Intrusion Detection System (IDS). An IDS system allows DOR to monitor traffic patterns and identify potential vulnerabilities for the network.

Research Investigators are provided with a desktop and/or laptop. In certain situations, investigators also have access to a local printer and mobile device. Computers are refreshed every three years. Standard software such as a secured email client, Microsoft Office 2010, and other software such as EndNote, SAS, or R are provided as a standard. Desktops and laptops are encrypted. Access to these devices requires a username and password. A Microsoft active directory domain is implemented for authentication. Complex passwords are required for access.

The DOR maintains a High-Performance Cluster that is available to researchers for computationally-intensive analyses, including analyses that use genome-wide data with a million or more data elements. This High-Performance Cluster was developed under the auspices of the DOR's Research Program on Genes, Environment and Health (RPGEH); the RPGEH is described further below. The DOR cluster is a Beowulf cluster with an Intel Xeon 12-core 48-GB head node and 16 compute nodes with dual hex-core Intel Xeon 2.53 GHz processors each. Each compute node is configured with 96 GB of RAM, bringing the cluster total to 268 processor cores and 1840 GB RAM. The cluster nodes are all equipped with 10-Gb network interfaces and can access a NFS storage server with 76 Terabytes of raw storage using both SAS and SATA disks set up in a RAID configuration. The RPGEH cluster runs the Redhat Linux OS, provisioned using the ROCKS+ cluster management framework, and uses the SGE resource scheduler for job submission and management. General purpose scientific and standard bioinformatics software tools and programming languages (C, Python, R, Perl, etc.) are installed on the cluster. In addition, software for genome-wide association studies such as impute2 and PLINK are also installed and available for use on the system.

THE DOR RESEARCH DATABASE

The DOR has developed and implemented a research database on an Oracle platform that serves as a repository for KPNC clinical and administrative databases from legacy and KP HealthConnect® (KPHC) data sources. Legacy databases are those that have been in use by KPNC that pre-date implementation of the KP electronic medical record (EMR); KPHC is the KP EMR. KPHC and several of the legacy databases that are incorporated into the DOR Research Database are described further below.

The DOR RDB aggregates into a single data warehouse the hundreds of separate regional data sources that researchers have typically accessed to extract data elements needed for specific research projects (**Table R1**). It conforms both legacy and KPHC data into standard database structures that are optimized for research query and retrieval. It uses a security model that meets all HIPAA, IRB and KP requirements. The database is now used by KPNC researchers to support virtually all research projects, including epidemiologicalstudies, health services research, and informatics projects. The database was developed in parallel with the deployment of KPHC within all KP medical facilities. It receives daily updates from KPHC's clinical databases, in addition to other KP clinical data sources not available through KPHC, including the legacy data systems described below. Because KP has had some form of electronic health record for almost 40 years, those historical records, which are preserved in legacy databases, have also been incorporated into the Research Database. This provides a unique opportunity for researchers to study secular trends of disease, as well as following the health and disease status of large cohorts over time.

VIRTUAL DATA WAREHOUSE (VDW)

The VDW is a collaborative effort of the Health Care Systems Research Network (HCSRN), a consortium of research groups in 20 health care organizations across the US and in Israel; participating members of the HCSRN are shown in **Figure R2**. Essentially all research organizations embedded within HCSRN health care systems maintain the VDW as a platform to facilitate research. The VDW is a series of dataset standards and automated processes that allow programs written at one HCSRN Site to be run against other VDW sites quickly and with minimal site-specific customization. The VDW is "virtual" in the sense that data remain at eachsite; the VDW is thus a federated database and not a multi-site physical database at a centralized data coordinating center. At the core of the VDW are a series of standardized file definitions. Content areas and data elements commonly required for research studies have been identified, and data dictionaries created for each of these content areas, specifying a common format for each of the elements – variable name, variable label, extended definition, code values, and value labels. Programmers at each participating research organization have mapped and transformed data elements from their local data systems into the standardized set of variable definitions, names, and codes, as well as onto standardized SAS file formats. The extract, transform, and load (ETL) procedures produce SAS datasets with a common structure at each HCSRN site. Each HCSRN member site creates their VDW data tables follow specific, standardized guidelines. These guidelines are adhered to and also reviewed regularly for accuracy and relevancy, and adjusted as necessary to adapt to the changes and challenges in the EMR and the health-research environments, including, but not limited to the addition of new variables and content areas. For KPNC, the DOR Strategic Programming Group, which consists of half a dozen programmers, identifies data gaps, provides guidance and consultation, and facilitates data problem resolution.

Data standardization for VDW development includes: specifying common variable names, labels, coding, and definitions; defining business rules to extract and convert variables stored locally to the common standards; testing standardized data for consistency and accuracy across member sites; standardizing methods by writing macros that are used across projects; and, teaching researchers and analysts how to use the VDW to guide construction of analysis files for approved research projects. As source clinical and administrative are updated constantly, the VDW tables are also refreshed on a regular basis – in the case of KPNC, most data tables are updated on a monthly basis.

With this common data model in place, programs written at one HMORN site can then be ported and run against other participating VDW sites' data quickly and with a minimum of site-specific customization. For example, an analyst can write an analytical program based upon the local VDW data, and then send that program to collaborating sites. The collaborative programmers need only change the file-reference statements to the location of their own data, run the program without additional alteration, and send the results back to the originating site. This enables each site to control and protect their data behind their own firewalls. Only the output, which is reviewed and screened for PHI, is returned.

This standardized approach to KP/DOR data increases overall study efficiency, as data from various disparate KP legacy sources are integrated with KPHC data into uniform datasets of very high quality. These

datasets are then made available for research through a single channel, the DOR VDW. The VDW thus leverages the unique capabilities of KPNC's longitudinal, clinical and electronic databases and facilitates the analysis of health-plan-specific issues. At KPNC, data may be accessed via SAS programs running SQL extracts of Oracle tables located on the M9000 server.

Table R1. Number of Variables and Observations in the DOR Virtual Data Warehouse, as of Nov 29, 2016

Governance	Data Table/Domain	Number of Observations	Total Variables	VDW Common Variables	DOR-specific Variables
HCSRN	Demographics	15,408,856	26	10	16
HCSRN	Enrollment	56,502,832	32	32	0
HCSRN	Utilization	567,282,851	52	21	31
HCSRN	Diagnoses	1,313,003,125	34	12	22
HCSRN	Procedures	1,130,476,157	43	16	27
HCSRN	Death	1,437,554	13	6	7
HCSRN	Cause of Death	4,415,460	11	6	5
HCSRN	Tumor Registry	503,493	247	168	79
HCSRN	Laboratory Results	1,056,054,109	36	30	6
HCSRN	Census Location	29,911,862	22	10	12
HCSRN	Census Demographics	741,414	142	108	34
HCSRN	Pharmacy, Outpatient	495,224,718	27	6	21
HCSRN	Ever NDC (drug codes)	52,713	62	17	45
HCSRN	Vital Signs	521,304,360	41	20	21
HCSRN	Social History	69,617,236	50	45	5
HCSRN	Language	8,273,609	4	4	0
CESR	Personal Health Record (PHR) Registration	4,370,756	6	6	0
CESR	PHR Proxy	1,306,290	5	5	0
CESR	PHR Messages	197,002,013	10	10	0
CESR	PHR Tests	218,812,242	5	5	0
CESR	PHR Activity	2,392,055,938	5	5	0
CESR	VDW_ORDER_MED	467,706,608	39	28	11
CESR	VDW_ORDER_DX_MED	169,331,944	9	4	5
CESR	VDW_CLARITY_MEDICATION	158,189	11	10	1
CESR	Infusion Meds, Administered	6,670,282	51	34	17
CESR	Infusion Meds, Dispensed	7,207,108	70	40	30
CESR	Infusion Meds, Planned	4,889,248	29	42	17
CESR	Pharmacy, Inpatient	170,941,832	28	28	0

VDW Datasets/tables have historically covered the following concept areas, and are shown schematically in **Figure R3**:

- **Enrollment:** Contains a record for every period of enrollment for each person enrolled in the health care plan. Each record represents a period of time during which the information on the included variables was true. Multiple records are typically necessary to represent changes over time in, say, insurance type(s), or completeness of data capture.
- **Demographics:** Birthdate, gender, race, and ethnicity (as available) for everyone enrolled or ever treated at the health plan.
- **Census Demographics:** Contains 2000 and 2010 Census information based on home address. In addition to a geocode, it contains census-based neighborhood area characteristics on education, income, housing, and race information.
- **Provider Specialty:** One record per provider code, it includes health plan providers, and may also include the most common providers outside of the health plan.
- **Utilization** (e.g., encounters, procedures, diagnoses): Documents encounters between health care providers and patients, including inpatient, outpatient, and emergency department encounters. Data are sourced from any or all of: integrated electronic medical record systems (EMR) (e.g., KPHC), legacy electronic data systems, or claims data. Each encounter can have any number of associated diagnoses and procedures. In general, the intention of the encounter file is to describe all significant interactions between patients and medical providers. The tables include: inpatient stays, emergency department visits, other outpatient hospital services such as same day surgeries, ambulatory visits, non-hospital residential stays such as at skilled nursing facilities, rehabilitation centers, nursing homes, overnight hospice facilities, overnight dialysis facilities and home health encounters.
- **Pharmacy:** Documents outpatient pharmacy fills. Can be sourced from either or both of internal electronic systems and claims data.
- **Laboratory Results:** Contains laboratory test and results information for a limited set of tests. Sourced from EMR or legacy internal systems.
- **Vital Signs:** The vital signs table includes various physiological measures taken by health professionals during the clinic visit. The traditional clinical vital signs include body temperature, pulse rate, blood pressure, and respiration. Because of HCSRN and DOR investigator interest, the VDW Vital Signs table also includes anthropometry (height, weight) and tobacco use.
- **Social History:** The social history table includes various behavioral measures taken by health professionals, either during the clinic visit or often over the telephone or via questionnaires. These measures may include the use of tobacco, alcohol and illegal drugs, as well as sexual behavior and contraceptive use, all of which carry substantial privacy concerns.
- **Tumor:** Incident cancers typically sourced from SEER data; in the case of the KPNC, these data come from our internal KPNCCR, which conforms to SEER and NAACCR standards.
- **Death:** Contains death date and cause of death information**.**

Since establishment in 2009 of the **KP Center for Effectiveness and Safety Research (CESR)**, KP members of the HCSRN have expanded VDW content to include **Prescription Medication Orders** – not just fills as previously maintained in the Pharmacy file; encounters through the online member interface, kp.org, known as the **Personal Health Record**; detailed data from **Infusion Medication** encounters, including treatment plans, dispensed medications, and administered medications; and **Inpatient Pharmacy** dispensings. Some CESR VDW content areas are also implemented in a subset of non-KP HCSRN institutions.

Thus, some data tables are maintained by all HCSRN consortium institutions, while a richer array of data domains are maintained at KP institutions including KPNC, under the auspices of the CESR.

In addition to maintaining SAS datasets as a version of the VDW, the DOR also maintains an instance of the VDW within the DOR Research Database in an Oracle database. The DOR VDW contains many local variables requested by DOR programmers and investigators, in addition to the standard HCSRN or CESR variables. A comparison of the DOR VDW with the HCSRN standard VDW data tables and numbers of variables is provided in **Table R1**. The number of observations contained in the DOR VDW tables, as of March, 2015, is also provided in this Table.

Overall, the VDW presents a rich data resource for research applications, with data harmonized across various internal data sources. In the case of KPNC, data in most Tables date back to 1996 or earlier. As these data are in tables using formats that are similar to VDW databases in other HCSRN settings, the ability to conduct

efficient collaborative research across these systems is enhanced substantially, and provides an outstanding resource for collaborative studies, both within KPNC alone and across the HCSRN.

KPNC CLINICAL AND ADMINISTRATIVE DATABASES IN THE KPNC RESEARCH DATABASE

KP HealthConnect®

In December 2004, Kaiser Permanente began implementing a new integrated electronic health record (EHR) system designed by the Epic Systems Corporation (Verona, WI) to automate its patient files and make documentation of care more efficient and complete. This expanded the provision of care beyond the traditional face-to-face doctor patient office visit and hospital admissions.

The standard Epic products have been customized extensively for the needs of KP, where it is now referred to as KP HealthConnect® (KPHC). The system replaced many of the core utilization and clinical documentation legacy applications, including ambulatory visit check-in, hospital-based utilization, and clinical documentation of diagnoses, orders for tests and procedures, and prescribed Inpatient medications. In addition to replacing these and other legacy functions, KPHC enhanced significantly the scope and detail of available data. Several other clinical and administrative functions, including radiology, labs, and other diagnostic testing, tracking of outside claims and referrals, and health plan membership enrollment, continue to use legacy applications.

Kaiser Permanente implemented KPHC in a multi-year staged manner, with different modules being deployed on a Region-by-Region and facility-by-facility basis. Every KPNC Facility is up and running with the following major Epic modules:

- **Ambulatory – Outpatient:**
  - EpicCare Ambulatory: Order entry, in-basket, clinical documentation, decision support
- **KP HealthConnect Online:**
  - Epic MyChart: Member chart, benefit coverage, co-pays, appointments, refills, health encyclopedia
- **Inpatient:**
  - EpicRX: IP Pharmacy Application
  - ED Manager: Emergency Department Utilization
  - EpicED: Emergency Department application
  - Epic ADT: Admission, discharge, transfer application
  - EpicCare Inpatient: Order entry, MAR, clinical documentation, decision support
  - OpTime: OR(Operating Room) management, scheduling, pref cards, materials and clinical documentation
- **Abstracting and Coding KPHC Hospital Billing:**
  - Health Information Management (HIMS): Abstraction of ICD9 diagnoses and procedures at discharge
- **Oncology**
  - Beacon: Infusion medication documentation including chemotherapy, order entry, treatment

The KPHC application systems store their transactional data in a real-time operational data store called “Chronicles.” While limited reporting is possible directly from Chronicles, its primary purpose is to support the operational KPHC electronic health record. Access to the data by researchers and other analytical use depends on a reporting data repository called “Clarity.”

Clarity consists of a large relational database hosted on the Teradata server platform. Clarity is updated nightly by a feed from the operational Chronicles data, so in general, it reflects clinical events with one-day latency. A single, consolidated Clarity database resides in each KP Region, holding only that Region’s data. For the most part, KP Regional versions of Clarity are structured consistently, although some customization has occurred. Clarity is an extremely large and complex set of relational data tables – over 34,000 at present – containing historical data dating back to each facility’s deployment date. Queries and reports generated from Clarity are as comprehensive as the front-end KPHC applications, and provide information on patient outcomes, clinical effectiveness, and quality. The Clarity databases thus provide a data base platform for health services research and invaluable data for epidemiologic studies that reflect actual patient care.

KPNC LEGACY DATA SYSTEMS

In addition to KPHC, data are available for research and operational purposes in legacy databases; many of these continue to be in operation. Many of these databases have been operational dating back to the mid-1990's or earlier, providing reasonably comprehensive electronic health data for KPNC clinical and administrative services that predate implementation of KPHC. Example databases that are available to DOR researchers and incorporated into the DOR Research Database are described here.

KP Foundation Systems (KPFS) Membership Database

The KPFS California Membership System supports establishing benefits, creating contracts, enrolling members, billing and reconciling dues, general ledger reporting, creating purchaser and member letters, and distributing membership data. KPFS is the system of record for all CA membership and eligibility information.

Benefits

The Benefits Application is a management application used to define, manage, maintain and support KP Health Plan Products for Members and Purchasers. Concepts include benefit services, which are medical services classifications and used as the lowest level of detail for defining benefits in Foundation Systems. A benefit offering is a collection of benefit services with associated co-payments and thresholds. Benefit offerings are assigned unique sequence numbers and are combined to form a benefit plan. A benefit category is a specific classification of benefit offerings containing services that are delivered in a specific operational area (e.g., Provider Office Visit (PROV), Mental Health Outpatient (MHOP), Rehabilitation (RHAB)). There are approximately 25 basic benefit categories per Plan (e.g., outpatient care, emergency room visits, hospitalizations, surgery, hospice, durable medical equipment, pharmacy, imaging, laboratory, etc.).

Patient Demographics Database (PATDEM)

This is a database of health plan member demographics. The database contains, but is not limited to, medical record number, patient's name, address, sex, date of birth, telephone number, and email address. It also contains information about specific member characteristics such as deafness or language preference. The database is updated when new information is available, often during appointment registration.

Appointment databases

The appointment databases offer online capabilities to schedule and register patients for outpatient encounters, and were implemented in 1987 in all facilities. The databases contain, but are not limited to, medical record number, appointment type, department (specialty), date, location and provider. It also contains registration information at the time of the visit, including confirmation that a scheduled visit was completed.

Admissions, Discharge and Transfer/Case Abstract System (ADT/CABS)

This database captured all inpatient hospitalizations occurring at Kaiser Permanente (KP) hospitals in California prior to implementation of HealthConnect ADT. The database contains, but is not limited to, medical record number, admission date, discharge date, admitting complaint, principal discharge and up to eleven additional diagnosis codes, principal and up to seven additional procedure codes, and discharge status. The codes for diagnoses and procedures used ICD-9-CM and Diagnostic-Related Groups (DRG) codes. The full diagnostic record in ADT/CABS was most often completed within one to two months following discharge.

Outpatient Summary Clinical Record (OSCR)

The OSCR databases were phased in during the mid-1990s to capture all outpatient encounters at KP hospitals, medical centers and medical offices and were retired due to completion of HealthConnect® implementation. The database incorporates data from over forty different optically-scannable medical-specialty-specific forms. The appropriate form was generated at time of registration and contains a check-off list for the most commonly-used diagnoses and procedures, as well as space for write-in diagnoses. These data contain, but are not limited to, medical record number, registration information and procedure and diagnosis codes for each outpatient encounter. Codes for diagnoses and procedures are based on ICD-9-CM and CPT4 codes.

Surgery

The Legacy KP Anesthesia and Surgery Information System (KASIS) application was the Surgery Scheduling and Case Tracking system for KP California Divisions. KASIS consisted of a module-based application from Surgical Information Systems (SIS) as well as two reporting products from Internal Quality (IQ), IQ Personal and IQ Objects. The modules available to all facilities include those related to scheduling and patient tracking, as well as a Preference Card Manager Module, Materials Manager Module, and administrative modules. Post-operative data-entry (PODE) sites have the Nursing Intraop module. The point-of-care sites

use the Nursing Intraop Module. Additional modules available to all sites included a Nursing Preop module, a Pre-Admission Testing module and a PACU module.

The KASIS data were held in ORO DB2 tables, which provide utilization information on surgical events that took place in Kaiser Permanente operating rooms. They contained data for the years 2002 and beyond. Prior to the year 2001, Northern California operating room utilization data were captured in a system called ORSOS, a PC-based computer system. Data from ORSOS were periodically uploaded to mainframe computer VSAM files, and then loaded into DB2 tables. In 2000, Kaiser converted from ORSOS to KASIS. KASIS data from both Southern and Northern California now reside in the same DB2 tables as the legacy ORSOS data. The KASIS data in DB2 are assigned to one of three different edit levels: Key (K), Utilization (U), and Complete (C). KASIS DB2 data includes utilization for both Northern and Southern California facilities.

A number of facilities used the KASIS system to track other events such as Labor & Delivery, Special Procedures, and Radiology, as well as surgical events performed at non-KFH hospitals, accounting for approximately seven to eight percent of all cases entered in KASIS.

Radiology (MRMS, TRRS)

The Medical Records Management System (MRMS) and Transcription Results Reporting System (TRRS) are closely related. TRRS contains in-house radiology procedure results, while MRMS provides the monthly data summary files intended for divisional accrual estimates and management reports; the latter aid in reporting, billing and analysis. MRMS is used in conjunction with TRRS. For example, the MRMS “encounter” file of visits will be joined with the TRRS “results” file to obtain the full scope of data used for a given report.

The files contain radiology visit records including X-Ray, MRI, and CT for KPNC, for both member and non-member patients. These systems were established in 2000.

Home Health / Hospice Services (HMS, HCMS)

These databases contain Home Health and Hospice program admissions. They also contain data describing patient services provided by KP agency employees, and some information on services provided by outside agencies (diverts). Data only contain service activity information that can be associated with a specific patient; employee activity information is outside the scope of these data. The current HCMS system was implemented in 2001 for KPNC facilities, with earlier data captured on the HMS system (1996-2000). The HMS data was folded into the HCMS. For Fresno, Modesto, Stockton and Stanislaus, HCMS was implemented in 2003. Prior to that time, activity was overseen by Outside Referrals, with data stored in AOMS. Some legacy HMS data elements were not included in the HCMS transition.

Authorized Outside Medical Services System (AOMS)

AOMS is a Kaiser Permanente database created in 1988, and contains data related to encounters and costs for medical services authorized by KP providers for health plan members, but carried out by non-KP providers. The AOMS system may also contain information on services provided by KP physicians in non-KP owned facilities. Examples of utilization included in AOMS are cardiac surgery, kidney dialysis, organ transplants, psychiatric hospitalizations, imaging and radiology, and skilled nursing facility admissions. The data in AOMS includes, but is not limited to, medical record number, referring KP physician, reason for referral, ICD-9-CM or CPT4 diagnostic and procedure codes, type and name of service vendor, dates of services, and amounts paid by KP for each service.

Skilled Nursing Facilities (SNF)

The Skilled Nursing Facilities (SNF) data provide both authorization and payment information for skilled and custodial stays at outside (non-Kaiser) Skilled Nursing Facilities for 1994 through the present. The source of SNF information is the Authorized Outside Medical Services (AOMS) referrals payment system. AOMS furnishes patient tracking data for authorized SNF days, by payment source and level of care. It also provides invoice information for cost analysis, as well as patient demographic information. These data are contained in SAS-formatted files on the KP mainframe platform. In 2006, Walnut Creek began documenting MD visits to outside skilled nursing facilities in KPHC Ambulatory clinical documentation software instead of Lotus Notes CMS. In August 2008, Providers from all facilities began entering their visits into KPHC.

Outside Medical Services Reporting (OMSR)

The OMSR system consolidates and simplifies AOMS inpatient hospitalization data, providing a single record for each authorized hospital stay. The OMSR data are stored in IBM DB2 format on the KP mainframe platform. Retrieval of data can be accomplished using SQL and transferred to other platforms.

Claims, Adjudication and Tracking Systems (CATS)

The CATS database contains encounter and cost data for services claimed for reimbursement by health plan members, but not authorized by KP providers. CATS data typically represent either claims for emergency treatment, or services provided outside the Northern California service area. The data in CATS include, but is not limited to, medical record number, reason for claim, ICD-9-CM or CPT4 diagnostic and procedure codes, type and name of service vendor, dates of services, and amounts paid by KP for each service.

Laboratory Utilization and Reporting System (LURS)

The laboratory database captures all ordered and performed laboratory tests from Kaiser Permanente hospitals, medical centers and medical offices. Created in 1994, it is part of a larger Region-wide Integrated Laboratory Information System (RILIS). The database contains, but is not limited to, medical record number, KP facility code, patient name, name of ordering provider, test or procedure name, results, date of test/procedure/result and abnormal or out-of-range flags. An online feature is the Patient Results Reporting System (PRRS). This is a real-time system for clinicians to retrieve accurately and rapidly results from the LURS database from their office or exam room without having to obtain the patient's medical chart.

Pathology (CoPathPlus, CPP)

CoPathPlus is the pathology accessioning/tracking/processing/diagnosing/resulting database. It is also known as APEX, and was rolled out in 2005. It contains detailed information on tissue specimens sent to KP pathology departments for preparation and evaluation. Some of the CoPathPlus information is transmitted to LURS (Laboratory Utilization and Reporting System), including the corresponding SNOMED (Systematized Nomenclature of Medicine) codes for anatomic pathology, while other elements, such as histology, are not. The precursor to CPP was RAPTOR, which in turn was preceded by RILIS.

Pharmacy Information Management System (PIMS)

The PIMS database contains prescription medications dispensed at Kaiser Permanente hospitals, medical centers and medical offices. The database, which was created in 1990 and implemented in all facilities in 1994, captures all data related to any prescriptions that were filled at any inpatient and outpatient KP pharmacy. The data in PIMS contains, but is not limited to, medical record number, cost, prescribing practitioner and medication information such as medicine name, National Drug Code (NDC), date of prescription, dosage and refill information. The database may be searched by NDC or a specific therapeutic class, as well as other variables mentioned above. This is a real-time system in that the labels used on dispensed drugs are generated by this system, and therefore the system is considered extremely accurate.

Ambulance

Ambulance data include information on the overall processing of Northern & Southern California ambulance data, and includes information about data elements that comprise the American Medical Response (AMR), Ambulance Transaction data (AMTN), Ambulance Utilization Review System (AURS), Claims Adjudication and Tracking System (CATS), Outside Care Processing System (OCPS), and Lotus Notes Ambulance Tracking and Ordering System (LATOS) SAS datasets. All datasets except Ambulance Transaction data (AMTN) are available starting in 1995 in a single combined library. The AMR data are available both at the Transaction level (AMTN) and in Summarized form (AMR). LATOS is Kaiser's own internal data for ambulances ordered; the data for Northern California trips are available in a separate library beginning in August 15, 2000 to date (i.e., the date of implementation of the AMR contract).

Cost Management Information System (CMIS)

CMIS is a decision-support system (DSS) that integrates health care utilization data with the General Accounting Ledger (GL) to provide fully-allocated costs by medical center, patient or service. The CMIS was created initially in 1994 and continues to be modified by incorporating additional data sources. CMIS utilizes various health care encounter systems and derives unit costs by allocating actual service department expenses to the weighted service volumes provided by the department. Service weights are developed internally to reflect KP operations. Overhead costs for administering the medical care program are allocated to unit costs via a step-down method. Indirect program costs equivalent to the costs of "insurance-related" functions, such as membership accounting and some regional health plan administration, are excluded. Utilization is accumulated for each patient's encounter with the health system, and the encounter is costed out by applying the service unit costs to the patient's actual utilization of services during that encounter.

KPNC DOR REGISTRIES

In addition to the various databases maintained by KPNC entities outside of the DOR that are listed above, the DOR maintains various registries that build upon or complement these data, and that are available for research or operational purposes. Select databases include.

Kaiser Permanente Northern California Cancer Registry (KPNCCR):

The KPNCCR is a database of cancer information on all patients who were diagnosed with a new cancer at a KP facility. The KPNCCR provides accurate data on clinical outcomes and cancer epidemiology and has the ability to identify cases for early intervention and the facilitation of research. The KPNCCR verifies key patient identifiers, consolidates data, links multiple primaries, performs follow-up, matches death certificates, and reports outcomes. The data in KPNCCR contains, but is not limited to, medical record number, date of diagnosis, primary site, histology, stage at diagnosis, date of last contact and vital status.

Maintained for regulatory and research purposes, the KPNCCR reports cancers no later than six months after diagnosis to the Northern California Surveillance, Epidemiology and End Results (SEER) Registry at the Cancer Prevention Institute of California (Fremont, CA), the Greater California SEER Registry at the Public Health Institute (Sacramento, CA), and the State of California Cancer Registry at the University of California Davis (Sacramento, CA). Data standards are thus consistent with those of the National Cancer Institute's SEER Program, and of those of the North American Association of Central Cancer Registries (NAACCR). The KPNCCR captures all cancers diagnosed or treated at KP facilities in the counties covered by the Northern California SEER Registry since 1973 and in all KPNC facilities since 1988, the years in which the respective regional and State cancer registries were implemented. It also captures cancers diagnosed at the KP Oakland Medical Center since its founding in 1942.

Starting in 2011 with the KP Santa Clara hospital, KPNC medical centers began to comply with standards set by the American College of Surgeons Commission on Cancer (COC). This includes following all cancers that are treated at that facility for recurrence. Thus, recurrences that occur in cancers treated at KPNC facilities that follow COC accreditation standards are also documented in the KPNCCR. The KPNC hospitals and years in which they began (or will begin) documentation of recurrence are provided in **Table R2**.

Table R2. KPNC Medical Centers and Years in Which Cancer Recurrence began to or will be Documented

Medical Center(s)	Year COC Standards Instituted/Planned	Year COC Accreditation Obtained/Planned
Santa Clara	2011	2013
Roseville, Sacramento, Walnut Creek	2013	2014
Antioch, Oakland, Richmond, San Francisco, San Jose, Vacaville, Vallejo	2015	Late 2016
Manteca, Modesto, Redwood City, South Sacramento	2016	Late 2017
Fremont, Fresno, San Leandro, San Rafael, Santa Rosa, South San Francisco	TBD (2016-2017)	TBD (2017-2018)

Kaiser Permanente Diabetes Registry of Northern California:

The Kaiser Permanente Diabetes Registry of Northern California is a database of KP members diagnosed with diabetes since 1971. Inclusion in the registry is based on specific criteria for each identifying source, and is based on a combination of factors such as diagnoses, laboratory values, and prescription medication use. The latest version of Diabetes Registry uses the DOR Virtual Data Warehouse (described below) to identify diabetics. The primary data sources feeding the VDW are Clarity data from HealthConnect®, Pharmacy Information Management System (PIMS), Region-wide Integrated Laboratory Information System (RILIS) / Laboratory Utilization and Reporting System (LURS), Outpatient Summary Clinical Record (OSCR), Admissions, Discharge and Transfer / Case Abstract System (ADT/CABS) and Authorized Outside Medical Services / Claims, Adjudication and Tracking System (AOMS/CATS). A small percentage of diabetics are identified through specific research surveys from 1993-1998. The Diabetes Registry data are contained in a SAS-formatted file on the KP mainframe platform.

Kaiser Permanente Mortality Linkage System:

The linked mortality database contains linked death certificate information for KP members who have died in California since 1966. Each year, all active and non-active KP members are linked to California state death certificates and the US Social Security Administration Death Master files using the following identifiers: social security number, name, date of birth, ethnicity, and place of residence. The linkage program assigns a probabilistic weight to each purported match (similar to the National Death Index), allowing users to choose how conservative they want to be in accepting matches as valid. The linkage file is updated annually approximately one year after the close of the calendar year in which deaths occur. The mortality file contains, but is not limited to, the following fields: medical record number, date of death, place of death, and ICD-9 coded underlying cause of death. The linked mortality data are contained in a SAS-formatted file on a mainframe platform, and in an Oracle database.

MAJOR RESEARCH RESOURCES IN THE DIVISION OF RESEARCH

Cancer Research Network

The Cancer Research Network (CRN) is a multi-site platform for cancer research established with NCI sponsorship in 1999 and last renewed from 2012-17 (U01 CA171524). KPNC's involvement in this project paves the way for more ambitious and systematic collaborations between the CRN and other NCI initiatives. The CRN includes 8 non-profit integrated health care systems plus 4 affiliate systems that provide comprehensive care to more than 10 million individuals, or approximately 3% of the U.S. population; all institutions are members of the Health Care Systems Research Network. While two-thirds of patients reside in California, the populations are diverse in geography and race/ethnicity. The scientific productivity of the network is demonstrated by the involvement of approximately 200 investigators at CRN institutions and collaborations with researchers at many other institutions, in more than 300 research grants with total costs of more than $40 million. The CRN includes several Scientific Working Groups that serve as natural binding sites for an array of research projects, not confined solely to cancer, including in Health Care Quality & Cost, Communications & Decision Making, Treatment & Outcomes, and Communication & Dissemination.

Comprehensive Clinical Research Unit

The Comprehensive Clinical Research Unit (CCRU) at the Division of Research was founded in 2008 to provide regional operational oversight for clinical trials and clinical research projects across KPNC Medical Centers, and to assist clinician researchers with all aspects of planning for and conducting research in conjunction with other regional entities. The CCRU team provides specific guidance to clinical trialists and clinician investigators on study design, project feasibility, budget development and evaluation, IRB preparation, staff planning, resourcing and training, trial/clinical research project implementation, collaborative efforts for monitoring of participant enrollment and quality for the project, and support for preparation for internal and external audits and other regulatory-related processes. The CCRU has developed a web-based portal to facilitate rapid requests for support services (Research Collaboration Web Portal, rcwp.kaiser.org).

Another critical part of the CCRU mission is to enhance the training of healthcare professionals and research staff, and it has developed a comprehensive online clinical research training program through KPLearn.org. This program comprises multiple modules (30-45 minutes each) related to key clinical research principles and pragmatic aspects of conducting research. Specific modules include: Choosing a Research Question; Data Sources; Confounding and Bias; Observational Study Designs; Experimental Designs; Study Subjects; Variables; Describing the Data; Hypothesis Testing and P-values; Sample Size and Power; and Multivariable Analyses.

Kaiser Permanente Research Program on Genes, Environment, and Health

The Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH) is a DOR resource designed to facilitate large-scale epidemiologic studies of genetic and environmental influences on common diseases and healthy aging. In addition to the research resources available through the DOR, the RPGEH provides approved researchers with access to (1) biospecimens collected from more than 200,000 adult members of KPNC, who have signed broad consents for the use of these specimens for research; (2) health survey data on 430,000 adult members, including those participants with biospecimens, which includes questions related to sociodemographic factors, tobacco and alcohol use, diet, health history, family history, physical activity, reproductive history, and more; and (3) environmental data, from existing public sources, which can include information on air pollution, water quality, pesticide use, green space, transportation, and

other factors, all of which can be linked to residential addresses of RPGEH participants. On a subset of 100,000 RPGEH participants, genome-wide genotypic data (> 675,000 SNPs) and telomere length measurements is available. These data and biospecimens can be linked with all of the administrative and clinical databases described above. Researchers can gain access to RPGEH resources and linked EMR data through application to a web portal: https://rpgehportal.kaiser.org/.

GROUP HEALTH RESEARCH INSTITUTE
FACILITIES & RESOURCES

Obesity, Chemotherapy Dosing, and Breast Cancer Outcomes

GHRI

Group Health Research Institute is an operating division of Group Health Cooperative, a nonprofit health system in Washington State. Group Health Cooperative has signed an agreement to be acquired by Kaiser Permanente, a national nonprofit health system. Subject to regulatory review and approval, and completion of other conditions to close, the proposed acquisition is expected to be complete in early 2017. If the acquisition proceeds, Group Health Research Institute will become part of the Kaiser Permanente research network. The transition is not expected to affect ongoing or proposed research projects.

Operating since 1947, Group Health and its subsidiaries provide medical coverage and care to approximately 630,000 members in Washington and North Idaho. In 2015, the latest year for which complete numbers are available, Group Health total annual revenue was $3.7 billion.

GHRI was established in 1983 with a mission to improve health and health care through high-quality, public-domain research, innovation, and dissemination. This includes epidemiologic, health services, and clinical research on the best preventive measures, screening tools, and interventions; the comparative effectiveness of different treatments; how health care delivery can be more efficient while improving health outcomes; and providing health care in practical, patient-centered ways. GHRI investigators focus on prevention and effective treatment of major health problems. Their research includes health behavior change; evaluating the efficacy and cost effectiveness of health services and technologies; population-based surveillance of health status in the Group Health population and beyond; and evaluating Group Health's programmatic decisions.

A key strength is the ability to study Group Health members over time. Access to this well-defined population provides unequalled opportunity for research and evaluation. Nearly all care for the stable Group Health population is provided within the system. Decades of high-quality, computerized data are available thanks to extensive automated data systems that track diagnoses and services such as laboratory, radiology, and pathology. Care provided outside Group Health is typically recorded in the claims database. Through their connection to Group Health, GHRI researchers gain crucial insights into the clinics, care providers, and communities from which the study data will come, and a relationship with the patients who may directly benefit from the results of their research.

GHRI, located in Seattle, has more than 300 employees including more than 60 faculty members, most with joint appointments at the University of Washington. GHRI includes the MacColl Center for Health Care Innovation, which develops, tests and disseminates strategies that improve health care quality; the Center for Community Health and Evaluation, which designs and evaluates health-promoting programs across the country; and the Group Health Department of Preventive Care, which is part of Group Health Physicians and provides a direct connection to its clinical practice group. In 2015, the latest year for which complete numbers are available, GHRI's total annual revenue was $48.4 million, with $34.6 million in grant funding from federal sources, primarily the National Institutes of Health. Substantial grant revenue also comes from private foundations such as the Robert Wood Johnson Foundation, and other agencies such as the Patient-Centered Outcomes Research Institute (PCORI).

CLINICAL

The Group Health family of organizations includes GHRI, Group Health Options, Inc. (1990), Group Health Foundation (1983), and Group Health Permanente medical group (1997)—with whom Group Health has an exclusive contract. In October 2005, Group Health also acquired KPS Health Plans, a health care service contractor in Northwest Washington.

The Group Health population: As of April 2016, Group Health's enrollment was about 630,000, including nearly 90,000 Medicare Advantage beneficiaries. In 2014, the Group Health Cooperative Clear Care® (HMO) Medicare Advantage plan was awarded a Medicare 5-Star rating for delivering quality and service to Medicare beneficiaries. It was one of only six plans in the country to receive the award in each of the previous three

years. Group Health provides primary, specialty, hospital, home health, and inpatient skilled nursing care on a prepaid (capitation) basis. Mental health and substance abuse services are part of the enrollee benefit package. Enrollees choose their primary care medical center and personal physician. Nearly two-thirds of members receive comprehensive care at Group Health Medical Centers. Group Health operates 25 primary care or family medical centers in 17 cities and several 24/7 urgent care centers. Group Health members have access to hundreds of specialists in 60 specialties and subspecialties including allergy and immunology; behavioral health; cancer care; endocrinology; eye care; heart care; hematology; neurology and neurosurgery; orthopedics; speech, language and learning; and women's health.

Group Health membership: The membership of Group Health reflects the surrounding community. As of July 2015, Group Health membership was 76% white, 10% Asian, 5% black or African American, 1% American Indian/Alaska Native, 1% Native Hawaiian or other Pacific Islander, 3% other, and 4% unknown (for members with self-reported race getting care at Group Health Medical Centers). Hispanic ethnicity is reported separately. Of Group Health members whose Hispanic/not Hispanic ethnicity was known, 5% were Hispanic. From 2010 United States Census data, the Washington State population is 77% white, 7% Asian, 4% black, 2% American Indian/Alaska Native, 0.6% Native Hawaiian or other Pacific Islander, and 11% Hispanic or Latino (5% reported two or more race/ethnicity categories).

Civil rights-era organizational policies aimed at preventing discrimination previously led Group Health not to collect race or ethnicity data. As federal regulations and quality-improvement perspectives shifted toward using such data to monitor care quality, Group Health began collecting this information from patients. Less than 1% of all Group Health enrollees choose to opt out of research.

Group Health has a slightly higher proportion of women (55%) than the regional community (50%) and the United States (51%). Members are also older (52% ≥45 years) than the regional community (38%) and the US (40%). Compared with the rest of the country, Group Health members are more likely to be Asian or Pacific Islanders (US: 6%), but less likely to be black (US: 14%) or report Hispanic ethnicity (US: 17%).

Group Health staff: Group Health Physicians is an American Medical Group Association Acclaim Award-winning medical group of 1,350 physicians and advance practice clinicians that includes family practitioners, behavioral health doctors, eye care providers, GHRI research faculty, and more than 700 specialists trained in more than 60 specialties and subspecialties. Most are board certified (96% family practice, 89% pediatrics, 94% Ob/Gyn, 94% other specialties). Group Health has affiliations with 39 other major institutions that meet its standards and follow its protocols. Group Health Cooperative also contracts with over 9,000 independent doctors and group practices throughout the state. These institutions provide care for Group Health members as permitted by the member's individual plan, when specialty services are unavailable at Group Health Medical Centers, or in locations where a Group Health Medical Center does not operate. In most cases, Group Health is billed for services received outside the system.

COMPUTER

An outstanding feature of Group Health as a study setting is the breadth and depth of its automated patient care and administrative databases and the highly developed information technology infrastructure that supports them.

Electronic health record system: The center of Group Health’s data systems is an EpicCare electronic health record (EHR) system, fully deployed at all Group Health-owned clinics and specialty centers as of November 2005. This full-featured EHR documents all health-related ambulatory care encounters in Group Health facilities, over the telephone, and by secure electronic messaging. Group Health uses Epic Corporation’s Clarity® reporting database to extract data from the EHR for administrative, research, and quality assurance purposes. The Clarity database is updated daily.

Group Health has implemented Beacon®, the Epic oncology module. Oncologists and pharmacists can use Beacon to manage laboratory and medication orders over the course of a patient’s care. Patient-specific treatment plans span encounters, allowing orders to be released in both ambulatory and inpatient settings. Beacon also tracks medication administration details.

For records before implementation of the EpicCare Ambulatory EHR, paper medical records are readily accessible. These records provide information from inpatient, outpatient, primary care and specialty care in a single comprehensive source. As an approved outpatient chart delivery site within Group Health's transportation system, GHRI maintains a computerized tracking and retrieval system and works with staff at Group Health's SeaTac Records Center to guarantee that paper outpatient records are accessible and secure. Previous GHRI records-based studies with active enrollees have retrieved 95% of paper records for all subjects.

Automated data sources for research: Electronic Group Health data sources include 1) enrollment and demographics, 2) diagnosis and procedures, 3) vital signs, 4) pharmacy/prescriptions, 5) laboratory, and 6) costs and claims.

1) Enrollment and demographics: Detailed enrollment and demographic data are available for all enrollees dating to the 1980s. This includes medical history number, age, gender, enrollment start and termination dates, type of coverage, primary care provider, and residential address and phone number. Enrollees can voluntarily report their race through the online health profile survey, and federal incentives are increasing the collection of demographic information including race, ethnicity and language in EHRs. Monthly snapshots of enrollment information and reasons for disenrollment have been recorded since 1988. Enrollment data are safely assumed complete because operating revenues depend on accurate tracking of enrollment status. Death certificate information derived from the Washington State vital records repository is updated annually for all people enrolled or receiving care at Group Health since 1972.

Geocoded census data are available for all health plan enrollees as of December 2011. Standardized measures of educational attainment, family income, and race summarized for census block groups are linked to each enrollee's residential address, providing a neighborhood-level look at useful socioeconomic measures.

2) Diagnoses and procedures: Diagnosis coding for patient encounters in Group Health-owned ambulatory care settings is guided by an interactive EHR interface that helps practitioners select appropriate International Classification of Disease codes (fully ICD-10 as of October 2015) and Current Procedural Terminology (CPT-4) codes. Procedures are coded at billing. Local modifications to the EHR's coding interface incorporate information previously made available to providers on paper forms.

Before the EHR, Group Health's treatment report forms (TRFs) guided diagnosis and procedure coding. TRFs were completed for each visit, except ancillary visits solely for lab, radiology, or physical therapy encounters. The forms were tailored to provide code sets relevant to different types of visits/care services. Providers checked off or wrote in up to 10 diagnosis (ICD) codes and up to 10 procedure (CPT-4) codes. Data from the TRFs were key-entered into the outpatient encounter tracking system. A 1992 audit found 97% of primary care visits identified in medical charts were represented in the tracking system. Claims forms are the sole source of diagnosis and procedure codes for encounters at facilities not owned by Group Health.

3) Vital signs: Group Health's EHR has captured clinically measured patient blood pressure, height, weight, and current smoking status during ambulatory encounters in all Group Health-owned clinics since November 2005. Some clinics have these data dating to October 2003. The system can capture multiple measures per person per visit. Vital signs data are available to researchers through Epic's Clarity® reporting database and the GHRI data warehouse.

4) Pharmacy/Prescriptions: Detailed data on prescriptions filled at Group Health's outpatient pharmacies have been maintained in a database since March 1977. Data include fill date, quantity dispensed, route, strength, cost, co-pay, and prescribing practitioner. Most prescription medicines are covered in full or with a modest co-payment (<$20). The pharmacy database is considered a complete source for data on prescription use. It has been used extensively for research, with previous studies showing that enrollees who receive care in Group Health-owned facilities fill nearly all their prescriptions at Group Health-owned pharmacies. For example, a survey of 990 postmenopausal women at Group Health found that 97% of subjects filled all or nearly all of their prescriptions at Group Health pharmacies. For prescriptions filled outside of Group Health but covered by Group Health insurance, we have pharmacy claims data.

5) Laboratory data: Data on all laboratory studies conducted at in-house labs for patients at Group Health-owned facilities are available dating back to 1988. Laboratory results from outside laboratories PAML (Pathology Associates Medical Laboratories) and INHS (Inland Northwest Health Services) are available since 2005. These data are stored in various databases, corresponding to laboratory data management systems used over the same period. The most commonly used lab data have been consolidated in standardized format in a unified data warehouse managed by GHRI containing 103.8 million records as of September 2012. Other lab data are available in separate data stores.

6) Cost and claims data: As of 1990, Group Health information systems capture and fully allocate costs of all internal services provided directly by Group Health as well as claims for covered services enrollees receive from contracted providers. Costs are allocated through a resource intensity weight assigned to each service, procedure, pharmacy fill or diagnostic test provided by Group Health or its contracted providers. Costs allocated to enrollees for services received from contracted providers are Group Health's payment to those providers. Although such services will more likely reflect market prices (services bought "on the margin") than services provided internally, they are the financial liability Group Health incurs in delivering health services to its enrollees. Both institutions and individual providers may submit claims for reimbursement. Detailed claims data, including information on diagnoses and procedures, are available dating back to 1979. The system can identify costs for specific encounters and services as well as aggregate costs for individuals over time.

Inpatient encounters in Group Health hospital are tracked through its hospital information systems, with data available dating to 1985. Services provided at non-Group Health facilities are captured retrospectively through a claims database. Starting in 1989, community-based encounters—including visiting nurse, hospice, respite, geriatric nurse practitioner services, nursing home rounds, and several other community-based services—have been tracked retrospectively in a computerized community health services system. This system captures provider, type of procedure, length of visit, up to 10 diagnosis codes, location of service, and price and billing information.

Enrollees may obtain care outside Group Health when they are referred for specialized services or have an urgent or emergency medical situation while out of the area, or by the design of an enrollee's health plan. As of 2012, contracted services represent approximately 50% of inpatient admissions and 50% of all specialty visits (or 15% of all ambulatory encounters) during a typical year. Costs to non-Group Health providers represent approximately 25% of total delivery system costs in a typical year. Approximately 50% of Group Health medical expenses are for care delivered by outside providers. Our claims data are believed to be complete since reimbursement for services provided outside Group health depends upon prompt submission of claims forms. The claims system database is widely used by many Group Health departments and receives a substantial amount of programming and maintenance resources—two indirect indicators of high-quality data. As claims forms are received and the data entered online, they are checked for valid CPT-4, ICD, and revenue codes. Online edits also check for codes appropriate to gender and age and for procedures appropriate to diagnosis. Internal audits of data take place at least semiannually. Internal audits in 2006 demonstrated 99% accuracy of claims payment and 98% accuracy of actual data input.

Cancer data

Surveillance Epidemiology and End Results Program (SEER)

At Group Health Research Institute, cancer incidence data come from the Seattle-Puget Sound Registry, which is part of the Surveillance, Epidemiology, and End Results Program (SEER) of the National Cancer Institute. SEER data are collected and maintained by the Cancer Surveillance System (CSS) project of the Epidemiology Program at Fred Hutchinson Cancer Research Center (Principal Investigator: Dr. Stephen M. Schwartz, Co-PI Dr. Christopher I. Li). CSS is currently funded by contract number HHSN261201300012I from SEER with additional support from Fred Hutchinson Cancer Research Center and the State of Washington. Since 1974, CSS has collected data on all newly diagnosed cancers (except non-melanoma skin tumors) in residents of 13 contiguous counties in northwest Washington State.

The SEER registries are considered the highest quality cancer registries in the United States. The Seattle-Puget Sound Registry has received the Outstanding SEER Registry Award and is consistently ranked among the top three in data quality among the 18 national SEER registries. Starting in December 2004, CSS standardized its data reporting to follow the North American Association of Central Cancer Registries format. New and existing

facility-based and central cancer registries use the same data dictionary to ensure that the definitions and codes used within their programs are standard and consistent with regional and national databases. The dictionary specifies data reportability, required and recommended data items, standardized item numbers and names, record layout, data codes, and coding rules. CSS transmits cumulative data on Group Health incident cancer cases monthly.

- Each month, GHRI uses ICD-9 and ICD-10 diagnosis codes and Healthcare Common Procedure Coding System codes to identify possible cancer cases and sends records to CSS for investigation. After review and abstraction, CSS transmits cumulative data on Group Health incident cancer cases monthly. These cancer data are stored in a SAS database on the Group Health Data Warehouse and are linked to other Group Health data by unique enrollee.
- From CSS: "Washington State law requires institutions and physicians to provide the CSS with information about each potential new cancer case within 45 days of diagnosis. CSS completes at least 95% of the case reporting for a given year by June 30th i.e., within 6 months, of the subsequent year. For example, as of June 30, 2010, the CSS had completed over 99% of the reporting for 2009 diagnoses."

Data Warehouses: Group Health's three data warehouses combine information from all relevant data systems to create a comprehensive, standardized, longitudinal record of each enrollee's encounter history. Two data warehouses are maintained by GHRI; Group Health maintains the third. Together, they provide rich, readily accessible data relevant to the most common types of research performed at GHRI.

1) GHRI Data Warehouse: GHRI has committed extensive resources over the past 10 years to building and maintaining its data warehouse, which contains data on more than 3.4 million former and current Group Health members. This research-centric data store was designed specifically to support the types of studies that GHRI investigators typically conduct. The GHRI data warehouse consists of a repository of SAS datasets and SQL Server databases covering all of the topical areas outlined above: enrollment, demographics, diagnoses and procedures, vital signs, pharmacy, lab tests, costs, and claims. A rich collection of powerful SAS macros can be used in conjunction with the GHRI data warehouse to generate commonly requested datasets based on user-specified criteria. GHRI data warehouse content and usage is documented by a wiki-based system authored by the data warehouse's architects, managers, and users.

2) Health Care Systems Research Network (HCSRN, formerly the HMORN or HMO Research Network) Virtual Data Warehouse (VDW): Originally created by the Cancer Research Network, the HCSRN VDW is an innovative data management system for multiorganizational, collaborative, data-based research. GHRI implemented the VDW at Group Health and continues to maintain it. The core of the VDW is a series of datasets with standardized structures, common data elements, and detailed documentation. The VDW is "virtual" because datasets are stored locally, at the 18 contributing HCSRN sites, rather than in a single, centralized database. Programmers transform data elements at their site onto standardized datasets using SAS software. Since the VDW files have a common structure, an SAS analyst at any HCSRN site can develop programs that can be run on data at all other sites. Standardized content areas within the VDW are enrollment, encounters, demographics, social history, tumor, pharmacy, diagnoses and procedures, census, vital signs, deaths and lab results.

3) Group Health Enterprise Data Warehouse: Created and maintained by Group Health IT staff, this data warehouse is available to all analysts at Group Health and GHRI. It provides source data for parts of the GHRI Data Warehouse as well as enrollment and clinical data in a variety of areas.

Computing Hardware/Software Infrastructure: GHRI has a highly developed information technology infrastructure. GHRI desktop computers run Microsoft Windows 7 and connect to file and print servers through a local area network (LAN) maintained by Group Health Information Technology Services. Additional resources include SQL database servers, Web applications servers running Windows 2008/2012, and a GHRI data warehouse. In-house technical support of GHRI's departmental servers and 300+ workstations is provided by the GHRI PC Support team. They also consult on technology purchases, set up new hardware and software, and provide one-on-one orientation to GHRI faculty and staff about the computing environment.

CONFIDENTIALITY

Confidentiality Procedures: All GHRI employees, including full- and part-time employees of the Survey Research Program, sign agreements to maintain confidentiality of data and research information. As a condition of employment, all members of the Group Health workforce must complete training in Health Insurance Portability and Accountability Act (HIPAA) requirements. All forms and study procedures are reviewed by Group Health's institutional review board (IRB) for compliance with HIPAA and human subjects protections.

Any data that will be accessed or disclosed outside Group Health will meet HIPAA requirements, through the use of business associate agreements, data use agreements, de-identification, and/or accounting of disclosures, as applicable.

All GHRI computers require passwords to access the network and the electronic mail system. Access to the Institute's data warehouse also requires special authorization. All paper (outpatient) medical records are delivered in secure bins to the Institute's confidential medical records room. Only authorized personnel are admitted, and the area is locked except during normal business hours. Staff who review medical records are trained in confidentiality procedures, including masking identifiers when photocopying, securing the records while at their desks, and returning all records at the end of their daily shifts.

GHRI policies and procedures ensure controlled access to computers and physical space for secure storage of data and confidentiality information. Access to GHRI's work areas is restricted by locked doors. Entry requires either a key card or a punch code that changes on a regular basis. Key cards or keys are required to enter the building, elevator, and GHRI floors during off-hours. Visitors must check in with designated staff to gain entry. A roster of persons authorized to enter the area is maintained by administrative personnel. Group Health requires employees to wear employee badges at all times and unfamiliar persons are required to state their purpose.

Health Insurance Portability and Accountability Resources: The Senior Manager of Information Technology, the head of clinical records for GHRI, and the Senior Manager of Research Human Subjects Review serve as "local experts" for review of HIPAA-related questions and issues. They provide consultations and recommendations for GHRI staff and the GHRI contract specialist.

OFFICE

GHRI occupies four floors (approximately 65,000 square feet) of the Metropolitan Park East building in downtown Seattle. Institute support staff includes groups of programmers, research project managers, and research specialists, each with more than 30 experts, all with the skills and experience required to execute large, complex research projects. The Institute also employs people with business skills and experience in administering federal and privately funded grants and contracts, generally about 15 grant management specialists and 20 administrative specialists.

Scientific and Support Staff

Eric B. Larson, MD, MPH, is GHRI's Executive Director. The GHRI workforce consists of approximately 300 full- and part-time individuals. Scientific staff include more than 60 faculty. Their graduate training and research experience span areas such as behavioral science, epidemiology, preventive medicine, family medicine, internal medicine, pediatrics, geriatrics, health services research, operations research, and survey research. Most GHRI investigators have faculty appointments at the University of Washington and contribute to its teaching and research programs.

Grant & Contract Administration

The Grant & Contract Administration (GCA) department of GHRI supports and advises researchers by reviewing, negotiating, approving, and providing administrative oversight related to grant and contracts administration. GCA acts in accordance with all applicable federal and state regulations, sponsor terms and conditions, and Group Health policies and procedures. GCA has grant management specialists trained in pre-award and post-award management of grants and contracts. A Contracts Services unit with legal, contractual and business experience is responsible for the legal and business aspects of contracts and agreements including compliance with industry and regulatory requirements. Work by GCA specialists includes preparing

and analyzing HIPAA data-use agreements (DUAs) to ensure that they outline appropriate data-sharing responsibilities while protecting patient privacy and data and specimen integrity. This oversight includes checking that DUAs incorporate privacy and confidentiality standards to ensure data security at the recipient site and prohibition of data manipulation for the purposes of identifying subjects. Contract Services represents GHRI in contractual negotiations with federal agencies, industry, healthcare and educational research institutions. All contractual reviews and approvals are completed by the Contracts Services unit before execution of any DUA by Institutional officials. If a situation requires expanded legal expertise, Group Health's legal department is staffed by practicing attorneys and available for review and consultation.

Affiliations

Group Health Medical Centers

Group Health Medical Centers are a network affiliate of the Seattle Cancer Care Alliance, a world-class cancer treatment center. Group Health also has affiliations with a variety of university and community college programs, including a formal affiliation with the University of Washington School of Public Health. In 2011, 269 medical trainees from 26 medical training institutions completed 25,800 learning hours in Group Health Medical Centers facilities. Nearly 200 Group Health Physicians medical staff in western Washington (27%) hold clinical appointments at the University of Washington or other institutions. Collectively, Group Health Physicians medical staff logged 7,010 teaching hours in 2011.

Group Health

The Group Health family of organizations includes Group Health Options, Inc. (1990), Group Health Foundation (1983), and Group Health Permanente medical group (1997)—with whom Group Health has an exclusive contract. In October 2005, Group Health acquired KPS Health Plans, a health care service contractor in Northwest Washington. Group Health is accredited by the National Committee for Quality Assurance at the highest level: "excellent." Group Health's hospital is accredited by the National Integrated Accreditation for Healthcare Organizations. The Joint Commission accredits Group Health laboratories and home care services with commendation.

GHRI

Cancer Research Group: The Cancer Researchers Group (CRG) at GHRI meets twice per month. Discussions focus on cancer research along the spectrum from prevention to end of life care. The CRG addresses all phases of cancer research projects, including data collection (such as Group Health's linkage with the Western Washington SEER cancer registry), study design, grant opportunities, and works in progress. Collaboration among cancer researchers at GHRI fosters synergy across research projects and creates efficiencies in data management and quality control. The CRG also has established relationships with cancer researchers from other local institutions and clinicians in the Group Health delivery system.

Breast Cancer Surveillance project: The Breast Cancer Surveillance project is a National Cancer Institute-funded initiative that collects data on risk factors, radiology screening and diagnoses (including all American College of Radiology-required data elements for mammography performance audits), pathology (via medical record abstraction), and cancer and vital status outcomes (via linkages to local registries). Data collected from this project are stored on the GHRI data warehouse and programmers regularly process the data for updates and quality control. The data are used for numerous research studies and for clinical purposes, including determining women's recommended screening intervals. The Surveillance team has extensive experience submitting secure, confidential data to the BCSC Statistical Coordinating Center.

Cancer Research Network: GHRI is a member of the Cancer Research Network (CRN), a consortium of research programs, enrolled populations and data systems of 14 non-profit research centers based in integrated health care delivery systems across the nation. All of these systems are members of the Health Care Systems Research Network (formerly the HMO Research Network))—a larger consortium of health maintenance organizations with formal recognized research capabilities and a shared commitment to public domain research. The CRN is funded by the National Cancer Institute through a cooperative agreement grant since 1999, which ensures substantive NCI involvement in attaining research goals and catalyzing new collaborations. In 2003, the Agency for Healthcare Research (AHRQ) joined NCI as a co-sponsor.

CRN research focuses on the characteristics of patients, clinicians, communities, and health systems that lead to the best possible outcomes in cancer prevention and care. The CRN allows for large, multicenter, multidisciplinary intervention research with the overall goal of increasing effectiveness of preventive, curative, and supportive interventions for common and rare malignancies. The research portfolio covers the full spectrum of cancer control, with studies on prevention, early detection, treatment, survivorship, surveillance, and end-of-life care. It comprises an infrastructure, core research projects, and a growing number of affiliated, supplemental, and pilot studies. The CRN also develops and uses standardized approaches to data collection, data management, and analysis across health through the virtual data warehouse. The virtual data warehouse organizes and documents data consistently across sites. It was designed to facilitate multisite research, but is also used extensively for single-site research, particularly at GHRI.

Collaborative expertise: GHRI is the lead site for several large, collaborative federally funded research projects, including several through the Cancer Research Network, a network of 19 health systems in the United States established in 1999. GHRI has also been instrumental in establishing several groups within the Health Care Systems Research Network (HCSRN, formerly the HMO Research Network) including the Cancer Research Network and the Mental Health Research Network.

GHRI is a collaborator on multisite federal research awards and contracts including the University of Washington's Clinical Translational Science Award (Institute of Translational Health Sciences [ITHS]), the Breast Cancer Surveillance Consortium (funded by the National Cancer Institute), the Vaccine Safety Datalink (from the Centers for Disease Control and Prevention), and the US Food and Drug Administration Sentinel initiative.

GHRI staff have extensive experience in monitoring and meeting the financial and logistical requirements of large programs projects (e.g., the CRN was funded for over $20 million for five years in its most recent grant cycle). GHRI also has extensive information technologies and communication support infrastructures, which will be leveraged for this project.

RUTGERS CANCER INSTITUTE OF NEW JERSEY (CINJ)

Dr. Elisa Bandera is a professor and co-leader of the Cancer Prevention and Control Program at CINJ. Established in 1990, CINJ is the first and only National Cancer Institute designated Comprehensive Cancer Center in NJ. Through its community-based partnerships, CINJ is striving to reduce the cancer mortality rate in New Jersey's underserved populations. CINJ is one of the research institutes of Rutgers Biomedical and Health Sciences (RBHS), now also hosting the New Jersey State Cancer Registry (formerly located at the NJ Department of Health and Senior Services). Supported by the NCI Cancer Center Support Grant, CINJ has established several shared resources, including Biospecimen Repository Service, Biometrics, and Office of Human Subjects Research.

Office: Dr. Bandera's office (room 5568, approximately 125 sq. ft.) and 6 additional cubicles for her research staff are located on the 5th floor of Rutgers Cancer Institute of New Jersey. In addition, the research team has access to several large conference rooms with teleconferencing capabilities.

Computers: Dr. Bandera's office and each cubicle are equipped with powerful computers and the necessary software (Microsoft Office, SAS, Adobe Pro, etc). Fax machines and network printing are also available.

EQUIPMENT

RUTGERS CANCER INSTITUTE OF NEW JERSEY (CINJ)

Major Equipment: CINJ Cancer Informatics Core Computational Resources has 2 SunFire X4450 servers each equipped with 4 Intel Xeon 2.93 GHz quad-core processors and 32 GB RAM for running research application software. Web server support for CINJ research faculty is provided in the form of 3 SunFire X4150 servers each equipped with 2 Intel Xeon 3.16 GHz quad core processors and 16 GB of RAM. For additional support, the Informatics core has 2 database servers (SunFire X4600: 1 production and 1 development) each with 4 dual core AMD Opteron processors and 16 GB of RAM. The database servers run Oracle, MySQL and SQL. A Sun StorageTek 2540 array (3 TB) and a Sun StorageTek J4200 (12 TB) array are both available for research storage and backup of data with weekly archive to a Sun SL500 tape library.

PHS 398 COVER PAGE SUPPLEMENT

OMB Number: 0925-0001

Expiration Date: 10/31/2018

1. Human Subjects Section

Clinical Trial? ❍ Yes ● No

*Agency-Defined Phase III Clinical Trial? ❍ Yes ❍ No

2. Vertebrate Animals Section

Are vertebrate animals euthanized? ❍ Yes ❍ No

If "Yes" to euthanasia

Is the method consistent with American Veterinary Medical Association (AVMA) guidelines?

❍ Yes ❍ No

If "No" to AVMA guidelines, describe method and proved scientific justification

3. *Program Income Section

*Is program income anticipated during the periods for which the grant support is requested?

❍ Yes ● No

If you checked "yes" above (indicating that program income is anticipated), then use the format below to reflect the amount and source(s). Otherwise, leave this section blank.

*Budget Period *Anticipated Amount ($) *Source(s)

4. Human Embryonic Stem Cells Section

*Does the proposed project involve human embryonic stem cells? ❍ Yes ● No

If the proposed project involves human embryonic stem cells, list below the registration number of the specific cell line(s) from the following list: http://grants.nih.gov/stem_cells/registry/current.htm. Or, if a specific stem cell line cannot be referenced at this time, please check the box indicating that one from the registry will be used:

Specific stem cell line cannot be referenced at this time. One from the registry will be used.

Cell Line(s) (Example: 0004):

5. Inventions and Patents Section (RENEWAL)

*Inventions and Patents: ❍ Yes ❍ No

If the answer is "Yes" then please answer the following:

*Previously Reported: ❍ Yes ❍ No

6. Change of Investigator / Change of Institution Section

❑ Change of Project Director / Principal Investigator

Name of former Project Director / Principal Investigator

Prefix:

*First Name:

Middle Name:

*Last Name:

Suffix:

❑ Change of Grantee Institution

*Name of former institution:

PHS 398 RESEARCH PLAN

OMB Number: 0925-0001
Expiration Date: 10/31/2018

Introduction	
1. Introduction to Application (Resubmission and Revision)	
Research Plan Section	
2. Specific Aims	Specific_Aims1014037292.pdf
3. Research Strategy*	Research_Strategy1014037291.pdf
4. Progress Report Publication List	
Human Subjects Section	
5. Protection of Human Subjects	Protection_of_Human_Subjects1014037290.pdf
6. Data Safety Monitoring Plan	
7. Inclusion of Women and Minorities	Inclusion_of_Women_and_Minorities1014037289.pdf
8. Inclusion of Children	Inclusion_of_Children1014037288.pdf
Other Research Plan Section	
9. Vertebrate Animals	
10. Select Agent Research	
11. Multiple PD/PI Leadership Plan	
12. Consortium/Contractual Arrangements	Consortium_Arrangement1014037036.pdf
13. Letters of Support	Letters_of_Support_Combined1014037157.pdf
14. Resource Sharing Plan(s)	
15. Authentication of Key Biological and/or Chemical Resources	Authentication_of_Key_Resources1014037162.pdf
Appendix	
16. Appendix	

SPECIFIC AIMS

Epidemiologic studies suggest that obese breast cancer patients experience poorer outcomes than their normal-weight counterparts (1-3), with a recent meta-analysis of 82 studies reporting that obese breast cancer cases have 41% higher total mortality than normal-weight cases (95% CI: 29%-53%) (1). It has been suggested that obese women may experience poorer outcomes, in part, due to the greater likelihood of dose reduction in chemotherapy (4-6). For most cytotoxic drugs, dose is calculated using body surface area (BSA); therefore, obese women would be expected to receive a higher absolute dose than normal weight women. However, due to concern about inducing chemotherapy-associated toxicity, clinicians are more likely to reduce the dose administered to obese women than normal-weight women (7), a finding replicated in several studies (8-12). While dose reductions may be warranted for reasons such as certaincomorbidities and toxicities, in 2012, the American Society of Clinical Oncology (ASCO) released guidelines urging clinicians to end this practice on the grounds of obesity, given research showing that obese women dosed at their BSA-determined dose are no more likely to experience toxicity than their normal weight counterparts (5). The guidelines were met with some criticism, citing the need for further research on adequate dosing (13).

Understanding the drivers of dose reductions may help better inform our understanding of this practice and efforts to disseminate guidelines; however, we know little about factors driving dose intensity, and how these factors may vary by body size.

Two major questions pertaining to the ASCO guidelines remain unanswered. First, while these guidelines suggest that chemotherapy dose reductions among obese patients may, in part, explain the association between obesity and breast cancer survival (5), this question has not been evaluated. **Demonstrating that dose reductions contribute to the associations between obesity and adverse outcomes would strengthen the case for these guidelines, and would provide a clear point of intervention to improve prognosis for obese women.** Second, the guidelines acknowledge that data pertaining to risk of toxicity are extremely limited with regard to *more severe obesity*, and in the real-world context of the presence of obesity-associated *comorbid conditions* (5). **If women with larger body sizes receiving the BSA-determined dose of chemotherapy do not experience excess toxicity as compared to normal weight women, this would provide further evidence in favor of the ASCO guidelines.**

Our interdisciplinary team, with expertise in epidemiology, pharmacology, and breast medical oncology is uniquely positioned to address these questions. Taking advantage of the rich data of two integrated healthcare delivery systems, Kaiser Permanente Northern California (KPNC) and Group Health (GH), in nearly 34,000 early-stage (Stage I-IIIA) breast cancer patients, we will address the following **Specific Aims:**

1. Identify predictors of chemotherapy dose intensity, focusing on whether body size is a principal driver of dose reduction, and whether other predictors of dose reduction vary by body mass index. Factors to be considered include patient-level factors (e.g., age, race/ethnicity, level of obesity, comorbidities such as diabetes, kidney disease, or cardiovascular disease), disease characteristics (e.g., stage, nodal involvement, grade, hormone receptor status), treatment, and provider-level factors (e.g., practice size, gender, age, provider-patient racial concordance).
2. Evaluate associations between body size and breast cancer recurrence and survival, focusing on the role of chemotherapy dosing as a mediator of these associations.
3. Among women identified as receiving BSA-expected dosing of chemotherapy, evaluate the association between BMI and toxicity; this will reveal if women receiving BSA-determined dosing at higher BMIs experience more or less toxicity than their non-obese counterparts. Toxicities include neutropenia and neuropathy, cardiotoxicity, renal impairment, and hepatic impairment.

Understanding these points will better inform clinical management for the estimated 102,000 obese women diagnosed with breast cancer each year in the United States (14, 15). It is critical that we address this issue, given that approximately 40% of adult women in the United States are obese and the prevalence of obesity continues to rise among women (14). With detailed information on treatment, comorbid conditions, and toxicities from KPNC and GH, as well as follow-up for recurrence and survival, we are uniquely positioned to address these questions about which little is known, but which have direct clinical relevance. In fact, to our knowledge, this is only such setting in which such a large-scale study can take place. Given our experience, expertise, and access to this rich data source, our team is ideally suited to address this important, novel, and timely avenue of research.

2. RESEARCH STRATEGY

2a. SIGNIFICANCE

2a1. OVERALL SCIENTIFIC PREMISE. Obesity has been associated with reduced survival and increased risk of recurrence among breast cancer patients (1-3). Obese women may experience poorer outcomes, in part, because clinicians may be more likely to depart from recommended chemotherapy dose levels when treating larger women. As dosing of most cytotoxic agents is determined by body surface area (BSA), a large BSA would result in higher chemotherapy doses, with the resultant fear that in heavier women, there may be a greater likelihood of chemotherapy-associated toxicity. Even so, in 2012, the American Society of Clinical Oncology (ASCO) released guidelines stating that obese women should be dosed according to their full body surface area (BSA), based on evidence that fully-dosed obese women do not appear to experience more toxicity than fully-dosed normal-weight women (5). Importantly, the ASCO guidelines acknowledge that data are extremely limited with regard to more severe obesity and in the real-world context of comorbidities. These guidelines were met with some criticism, citing the need for stronger evidence (13). Addressing the gaps in knowledge related to chemotherapy dosing in obese women will provide evidence to better inform clinicians and improve clinical care. Given the high and increasing prevalence of obesity among women in the US (16) and the large number of women diagnosed with breast cancer each year (15, 17), it is critical that we improve our understanding of optimal dosing of chemotherapeutic agents among obese breast cancer patients. Understanding these points can improve care for the estimated 102,000 obese women diagnosed with invasive breast cancer each year in the US (14, 15).

2a2. OBESITY AND BREAST CANCER OUTCOMES. Obese women experience worse breast cancer outcomes compared to normal-weight women. A recent meta-analysis of 82 studies reported that obese breast cancer cases have 41% higher total mortality than normal-weight cases (95% CI: 29%-53%), and 35% higher breast cancer mortality (95% CI: 24%-47%) (1). Compared to normal-weight women, obese women also experience increased risk of recurrence (2, 3). For example, in a recent retrospective cohort study in a screened population, obese women experienced a 2.43-fold (95% CI: 2.34-4.41) higher risk of recurrence than normal-weight women (2). While there is no consensus regarding the biological mechanisms driving these associations, it has been suggested that estradiol, insulin, and pro-inflammatory pathways may be involved (4, 6, 18). From a clinical practice perspective, it has also been suggested that inadequate dosing of chemotherapeutic agents among obese women may also contribute to poorer survival observed in this group (4-6). Specifically, if obese women are under-dosed due to concerns about inducing toxicities, they may receive less-than-adequate therapeutic dose, resulting in poor outcomes. However, most studies evaluating the impact of obesity on breast cancer survival have lacked sufficient treatment information and/or follow-up for adverse outcomes to assess this possible explanation.

2a3. OBESITY AND DOSING OF CHEMOTHERAPEUTIC AGENTS. The recommended dose of most cytotoxic agents is determined by BSA, and therefore obese women are expected to receive a higher absolute dose than their normal-weight counterparts. However, due to concerns regarding chemotherapy-associated toxicity, some physicians scale back the dose administered to obese women. Such dose reductions may take the form of dosing according to ideal body weight or the use of a BSA cap. A 2005 study among 9,672 women receiving care from oncology centers across the United States reported that 20% of women with class I obesity (body mass index, BMI 30-34.9 kg/m^2) and 37% of women with class II+ obesity (BMI≥35 kg/m^2) received less than 90% of their expected BSA-determined dose of the chemotherapeutic agents doxorubicin and cyclophosphamide in the first cycle of chemotherapy, while 9% of normal-weight women were comparably dose reduced (7). This association has been replicated in other studies, and obese women are more likely to experience dose reductions for other common chemotherapeutic agents, including docetaxel (8-12).

2a4. OTHER DETERMINANTS OF DOSE REDUCTIONS. Other studiessuggests that older age, African-American race, low neighborhood education, and presence of serious comorbidities may be associated with dose reductions, although obesity appears to be a strong predictor of dose reductions, independently of thesefactors (7, 8, 12, 19-21). However, as prior studies have examined a narrow set of potential predictors, the determinants of dose reductions remain poorly understood, and it is unknown if provider characteristics, such as clinician gender, age, or provider-patient racial concordance may be associated with dose reductions. It is also possible that in the presence of obesity, the factors driving dose reductions among obese women may vary from those in the general

population. Understanding these drivers, and how they vary by body size, is critical in order to better address this practice.

2a5. OBESITY AND TOXICITY. Study results suggest that, across cancer sites, larger patients receiving full BSA-determined doses of chemotherapeutic agents do not appear to experience excess hematologic and non-hematologic toxicity, as compared to fully-dosed leaner individuals (5, 7, 10, 22-29). In a retrospective analysis of the Cancer and Leukemia Group B (CALGB) study 8541, obese breast cancer patients (defined as those having a BMI ≥27.3 kg/m^2) receiving full BSA-determined doses of cyclophosphamide, doxorubicin, and fluorouracil (CAF) experienced no excess grade 3+ hematologic or non-hematologic toxicity during the first cycle, as compared to their non-obese counterparts (<27.3 kg/m^2) (22). Another observational study of 8,022 patients receiving doxorubicin (Adriamycin) + cyclophosphamide (Cytoxan) (AC) found that among those receiving full BSA-determined doses, the risk of febrile neutropenia decreased with increasing BMI, suggesting that, if anything, fully-dosed obese women have a lower risk of toxicity (7). Similarly, data indicate that among 368 women receiving cyclophosphamide, methotrexate, and 5-fluororicil (CMF), increasing BMI was associated with increasing leukocyte nadir values, and that this association does not appear to be driven by BMI-associated dosing differences (27). Notably, a number of these studies categorized high BMI at thresholds lower than what we use to define obesity today, pertain to other cancer sites, and/or are limited to a narrow set of regimens. More importantly, data pertaining to more severe obesity are limited, as are data among women with comorbidities, given that such women are underrepresented in clinical trials, as presence of comorbid conditions is often an exclusion criterion for such studies (5). Only a few studies have explicitly presented data on the impact of fully-dosed chemotherapy on toxicity among severely-obese women, and these studies have also been limited with regard to small sample sizes, narrow range of toxicities examined, limited regimens represented, and/or cancer sites (7, 29). Furthermore, a recent study of dose-dense therapy revealed that 173 obese women receiving full BSA-determined doses of chemotherapy experienced *more* toxicity than 2,435 fully-dosed normal-weight women (30). While the reasons for this finding are unclear and may be multifactorial (perhaps differential representation of BMI within the obese group or comorbid conditions), it is possible that the association differs for this study as it is the only study to evaluate dosing in relation to toxicity for dose-dense therapy, a relatively new but now common way of administering chemotherapy. Given this finding, it is therefore critical that we better understand the relationship between body size and toxicity in the context of modern therapies, in addition to better understanding whether women with larger body sizes receiving BSA-determined dosing experience excess toxicity as compared to normal weight women. Understanding these points across a broad spectrum of body sizes and treatments and in the real-world context of comorbidity will address clinician concern regarding toxicity, and can reduce persisting uncertainty regarding dosing.

2a6. DOSE-REDUCED CHEMOTHERAPY AND BREAST CANCER OUTCOMES. Research demonstrates that chemotherapy dose reductions negatively impact patient outcomes (5, 22, 31-35). For example, in the trial, Cancer and Leukemia Group B study 8541, it was found that the 519 patients randomized to the high-dose arm of cyclophosphamide, doxorubicin, and fluorouracil (CAF) experienced markedly better overall and disease-free survival compared to the 518 patients randomized to receive 50% of the dose received by the high-dose arm (35). Another study found that among ER- breast cancer cases, 596 women receiving less than 85% of the BSA-determined dose of CMF experienced significantly poorer 10-year overall and disease-free survival compared with 143 women who received ≥85% of the BSA-determined dose, with the same pattern observed in the obese group (31).

2a7. DOSING AS A POTENTIAL EXPLANATION FOR POOR OBESITY-ASSOCIATED OUTCOMES. The association between obesity and survival appears to be qualitatively stronger in observational studies than in randomized controlled trials (RCTs), which could be attributed, in part, to less dosing variability in trials (6, 36). This may also be attributed to strict eligibility criteria for trials, which typically exclude those with severe comorbidities, thereby limiting the inclusion of severely-obese women. Despite the suggestion that dose reductions may mediate the association between obesity and poor breast cancer outcomes, to date, there is no empirical evidence addressing if, and to what extent, dose-reduced chemotherapy mediates the association between obesity and breast cancer outcomes.

2a8. ADDRESSING GAPS TO IMPROVE CARE. Addressing these gaps—specifically, addressing unanswered questions regarding i) whether the risk of toxicity among fully-dosed women varies by body size and ii) the extent to which dose reductions may mediate the association between obesity and poor outcomes—can provide the evidence needed to better inform clinician decisions regarding optimal chemotherapy dosing. By addressing these gaps, we can provide clinicians with the evidence needed to address persisting uncertainty about optimal dosing of obese women and better informdosing practices. Furthermore, by understanding the predictors of chemotherapy dose reductions, particularly among obese breast cancer patients, we will gain knowledge about the drivers of this practice, and will be better positioned to disseminate findings that may impact clinical practice. With an estimated 102,000 obese women diagnosed with invasive breast cancer each year in the US, and the prevalence of obesity continuing torise among women, the proposed project can help address the obesity-associated disparity in breast cancer outcomes (1-3, 14, 15).

2b. INNOVATION. In this study, we will leverage the uniquely rich clinical data resources of integrated healthcare delivery systems to study the experiences of nearly 34,000 breast cancer patients. Given that the literature indicates a) that obese women are more likely to be dose reduced and b) that dose reductions may negatively impact long-term outcomes, there is an urgent need to assess the impact of dose reductions in chemotherapy on the association between obesity and breast cancer outcomes. By applying recently-developed epidemiologic methods in mediation analysis (37, 38), we can estimate what the association between obesity and breast cancer survival would have been if all women had received their full BSA-determined dose, a novel question that has never been addressed. Furthermore, given the large sample size, we will be able to evaluate the association between a range of BMI categories and toxicity. If results indicate that fully-dosed chemotherapy does not result in excess toxicity even at the extremes of obesity and in the presence of comorbidity, this may alleviate clinician concerns about inducing chemotherapy-associated toxicity at the full-BSA determined dose. This line of research is ideally approached using data from a large integrated healthcare delivery system, given the rich treatment data and longitudinal follow-up for recurrence and survival in the real-world setting (representing a broad spectrum of BMIs, treatments, and comorbidities). In fact, to our knowledge, this is only such setting in which such a large-scale study can take place. For example, other large data sources, such as the Surveillance, Epidemiology, and End Results (SEER)-Medicare linked databases, do not have data on chemotherapy dose or BMI; studies within oncology clinics or academic health centers typically do not have data on long-term health outcomes; and clinical trials are not only small, but are also limited by homogeneity of dosing and underrepresentation of those with comorbid conditions and severely-obese individuals. Furthermore, given the richness of the clinical data in this observational setting, we will be able to carefully control for and stratify by comorbidities. We therefore propose to address this question in a large cohort of breast cancer patients diagnosed and treated in two integrated healthcare delivery systems that are part of the Cancer Research Network (CRN, U24 CA171524, PI: Kushi). By bringing all of these variables together in a single, large study, we will be uniquely situated to address these important and timely clinical questions.

2c. APPROACH

2c1. RESEARCH DESIGN AND METHODS

2c1.1. Overview. The setting for this project is two integrated health care delivery systems, Kaiser Permanente Northern California (KPNC) and Group Health (GH). These health care systems have longstanding clinical and administrative databases that are available for research purposes, high levels of retention enabling capture of much of the clinical experience, and a long history of collaboration with each other and researchers external to these systems. As of February, 2017, GH became a part of Kaiser

Permanente, and is now known as Kaiser Permanente of Washington; for the purposes of thisapplication, we continue to refer to GH. GH's patientpopulation, providers, and research will remain fairlyconstant through this transition and the acquisition should have no impact on the proposed study.

2c1.2. Study Population. This study will be conducted using data on approximately 33,908 Stage I-IIIA female breast cancer patients, age ≥18y, diagnosed and treated at KPNC and GH. Detailed descriptions of these integrated healthcaredelivery systems are provided below.

KPNC and GH are members of the Cancer Research Network (CRN), a consortium of integrated health care delivery systems (39). At all CRN sites, electronic data are stored in a Virtual Data Warehouse (VDW), a standardized data modelwith data derived from clinical and administrative databases (40). The standardized data model provides consistent variable definitions and formats

across multiple sources within each site, facilitating harmonization across sites (41-45). Each site maintains its own data, yet data are easily shared across sites using standard naming and coding. Each site stores data in domains, including neighborhood/contextual data from the Census, demographics, enrollment, encounters, providers, diagnoses, procedures, laboratory results, pharmacy, infusion medications, tumor registry, vital signs, social history, and death. For example, the 'infusion medications' domain includes data on chemotherapy, including administered drugs, dates of administration, and dosages. For most domains, data are updated on a monthly basis, and a data dictionary is available on a password-protected website maintained at each site. In this study, relevant data will be pulled from VDW tables or other clinical and administrative data when needed (**Table 3**). Programmers at one site (GH or KPNC) will create the initial programs for such data pulls, and this program will be shared with the other site. Data will be compiled at the originating site and checked for consistency and quality control. A harmonized data set containing both sites' data will then be shared with Memorial Sloan Kettering Cancer Center (MSKCC) for analysis.

Table 3. Key variables, by data source

Variables	Data Source	
	KPNC	GH
BMI	VDW vital signs table	VDW vital signs table + medical charts review
Chemotherapy	VDW infusion medication tables	VDW tumor registry table VDW procedures table VDW pharmacy table + medical chart review (dose)
Toxicities	Various VDW tables (diagnoses, procedures, pharmacy, etc.)	Various VDW tables (diagnoses, procedures, pharmacy, etc.)
Recurrence	Algorithm of clinical and administrative data from VDW + medical charts	Algorithm of clinical and administrative data from VDW+ medical charts
Death/ cause of death	VDW death, cause of death tables	VDW death, cause of death tables

ABBREVIATIONS: BMI (body mass index); VDW (Virtual Data Warehouse)

2c1.2a. Overview of KPNC & Collection of Data at this Site. KPNC is one of the oldest and largest integrated healthcare delivery systems in the US. KPNC provides care to about 3.9 million members throughout the San Francisco Bay Area and the Central Valley of California, across a network of 21 hospitals and over 200 outpatient clinics. In 2005, KPNC adopted a completely electronic health record based on the Epic Systems (Verona, WI) platform.

This study will include stage I-IIIA cases diagnosed between 2006 and 2019 at KPNC. We will include those patients with at least one year

of enrollment prior to diagnosis and available body mass index (BMI) data, resulting in an estimated sample size of 28,546. Data collection will begin in 2006, as 2005 is the earliest year for which electronic data are reliably available, and this allows documentation of comorbid pre-existing conditions in the year prior to diagnosis. Cases will be identified from the VDW Tumor Table, which is sourced from the KPNC Cancer Registry (KPNCCR). The KPNCCR captures all cancers diagnosed or treated at KPNC facilities. The KPNCCR reports to the Greater California and San Francisco Bay Area cancer registries, both of which are part of the SEER Program; all data elements are consistent with SEER standards. Data will be collected on cases from the year prior to diagnosis (to capture pre-diagnostic comorbidities), and all patients will be followedfor toxicity, recurrence, and survival through 2019. Most of the data needed for the proposed study can be abstracted from the VDW. For example, as noted in Section 2a.2, data on dosing of chemotherapy drugs has been previously abstracted from the VDW for our preliminary project, demonstrating the feasibility of using these data for research.

2c1.2b. Overview of Group Health & Collection of Data at this Site. GH is an integrated healthcare delivery system providing care to ~620,000 members in Washington State and parts of Idaho. A large portion of GH's service area is located within the Western Washington Cancer Surveillance System region, a member of the SEER Program.

Data from GH will be collected by building on the existing infrastructure of the COmmonly used Medications and Breast Cancer Outcomes (COMBO) study, a prospective cohort study of 4,216 breast cancer patients (PI: Boudreau; R01CA120562) (46). Briefly, women were selected for inclusion in COMBO if they were: 1) age ≥18 y; 2) resided within the 13 counties covered by the Western Washington SEER Program; 3) diagnosed with a histologically-confirmed first primary stage I or II breast cancer between 1/1/1990 and 12/31/2008; 4) did not have bilateral disease; and 5) enrolled in GH's integrated group practice for 1+ years before and after breast cancer diagnosis (unless they died). Women were required to be alive and recurrence-free for 120 days after completing surgery. In this study, data were collected from the period 1 year prior to diagnosis onward (to document pre-existing comorbidities) using health plan administrative databases (including demographics, enrollment, inpatient and outpatient diagnoses and procedures, breast services and results, pharmacy dispensings, laboratory results, vital signs, and death), medical record review (paper and electronic), and specifics of cancer diagnosis from the SEER Program. Medical records were reviewed by trained abstractors who collected data variables such as toxicities and recurrence. Inter- and intra-rater reliability tests revealed excellent agreement for key variables such as recurrence (0.93).

For this particular study, we will exclude COMBO participants diagnosed between 1990 and 1993, given that electronic data on comorbidity and toxicity are not routinely available prior to this time in the VDW (precluding the consistent documentation of toxicity and comorbidity across data sources). We will therefore limit this study to the 3,332 COMBO participants diagnosed between 1994 and 2008. We will expand data abstraction to include patients diagnosed with Stage IIIA cancer between 1994 and 2008 (n=343), as well as those diagnosed with Stage I-IIIA cancer in 2009−2015 (n=1,687), otherwise maintaining the same inclusion criteria listed above. Data will be abstracted using the same abstraction tools as the original COMBO study, leveraging the VDW to pull relevant data where applicable. While weight at diagnosis was collected in the initial COMBO study, the timeframe of ascertainment was not as precise as is necessary for this particular study, and more importantly, no data on chemotherapy dosage were abstracted (nor are chemotherapy data available in the GH VDW until relatively recently). Therefore, on all GH participants, we will extract data from medical records on height and weight at the start of chemotherapy, as well as doses of chemotherapy drugs received, a process we have piloted to ensure feasibility and appropriate budget estimates. Analyses will be limited to those with available BMI data. Follow-up for toxicity, recurrence, and survival will be extended through 2019 using a combination of data linkages, algorithms, and chart abstraction (see section 2c4.4 for detail). In total, this study will include approximately 5,362 Stage I-IIIA breast cancer patients diagnosed and treated between 2004 and 2015 at GH.

2c1.2c. Summary of Data Sources and Data Collection. The data sources and data collection described in the sections above are summarized in **Figure 1**. The Division of Research at KPNC and the Group Health Research Institute (GHRI) have long histories of conducting collaborative research, and are well-equipped for the proposed project.

2c3. AIM 1. Identify predictors of chemotherapy dose reduction, focusing on whether body size is a

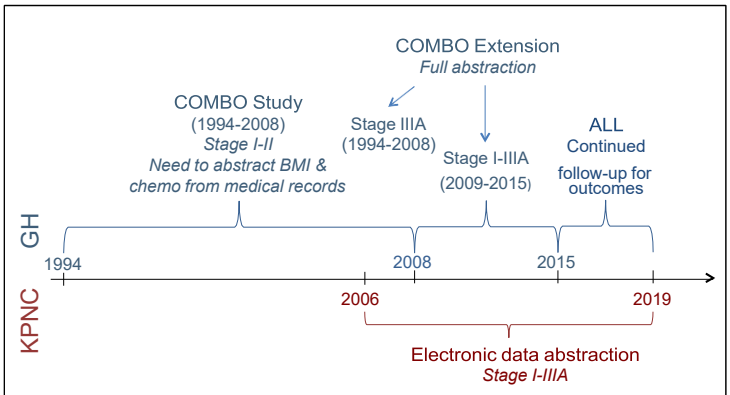

Figure 1. Overview of study sites and data abstraction at each site

principal driver of dose reduction, and whether other predictors of dose reduction vary by body mass index. These predictors include patient-level factors, provider-level factors, and disease characteristics,among others.

2c3.1. AIM 1. This question will be examined among early-stage (Stage I-IIIA) breast cancer patients receiving first-line adjuvant chemotherapy at KPNC or GH (n=11,325). All participants will be enrolled at the time of diagnosis and have received at least one cycle of adjuvant first-line chemotherapy. Underweight women will be excluded given the small number of underweight women receiving chemotherapy (n=170) and likely heterogeneity with the normal weight group (precluding collapsing of groups), leaving an estimated 11,155 women for analysis.

2c3.2. AIM 1. Exposure. We will first examine the association between obesity and dose reductions, and will further examine the factors associated with dose reductions within BMI strata, including patient-level factors, and provider-level factors, among others. Patient-level factors include age at diagnosis (quintiles), race/ethnicity (defined as non-Hispanic white, non-Hispanic black, Asian, Hispanic, other), level of obesity (measured by BMI (kg/m^2): class I obesity: BMI- 30-<35; class II obesity: BMI- 35-<40; class III obesity: BMI-40+), and comorbidity (history of diabetes; history of cardiovascular disease; Charlson Comorbidity Index). Provider-level factors will include practice size (number of oncologists: quintiles); practice location (urbanicity; quintiles); provider gender (male vs female); provider age (quintiles); provider-patient racial concordance (yes vs no). We will also examine how disease characteristics relate to dose reductions, including: stage (categorical: I, II, IIIA), nodal involvement (no vs yes), tumor size (<2 vs ≥2 cm), and estrogen receptor status (ER+ vs ER-). We will also examine how treatment relates to dose reductions, by cytotoxic drug and prior surgery (lumpectomy, mastectomy). Lastly, we will leverage the variation in study period covered by the two sites to capture how dosing practices have changed over this 26-year period. Note that in analyses of dose reduction over the course of therapy, toxicity will also be evaluated as a predictor of dose reduction. Information on provider characteristics will not be available from GH, and thus analyses of provider characteristics will be restricted to KPNC.

2c3.3. AIM 1. Outcome. Our primary outcomes in this analysis will be the *first cycle dose proportion*, calculated as the proportion of dose received relative to that expected. Secondary analyses will focus on the *average relative dose intensity* (ARDI) received over the course of therapy (defined as the average of the relative dose intensity (RDI); conceptually, this represents the dose proportion across all cycles, for each drug received). Understanding factors associated with dosing in the first cycle is of primary interest, as this reflects clinician intent (7) without also being affected by subsequent toxicity and treatment delays.

To calculate the first cycle dose proportion, the observed dose is divided by the expected dose. The expected dose is calculated by multiplying the participants' BSA (m^2) – a function of weight and height calculated using the Mosteller formula (47)) – by the prescribed dose (mg/m^2), as determined by calendar-time appropriate National Comprehensive Cancer Network (NCCN) guidelines (www.nccn.gov) for the given regimen. For example, if cyclophosphamide is dosed at 600 mg/m^2, a participant with a BSA of 2 m^2 would be expected to receive 2 x 600 = 1200 mg of cyclophosphamide. If she is observed to have received 900 mg, then she received 900/1200 = 75% of the intended dose. The value for each drug received in the first cycle is then averaged.

To calculate the dose intensity received over the course of therapy, we will calculate the RDI for each drug received, as shown in **Figure 2**. Calculation of the RDI, a standard measure in studies of chemotherapy dosing (7, 19, 20, 48), requires knowledge of the doses and intervals at which each drug was received (numerator), as well as the planned doses and intervals (denominator). By abstracting information on dates of administration and dosages, we can determine the observed dose and duration. Determining the expected cumulative planned dose and planned treatment duration is more complex.

$$\text{RDI} = 100 \times \left[\frac{\text{Cumulative dose (mg)} / \text{Treatment duration (wks)}}{\text{Cumulative planned dose (mg)} / \text{Planned Treatment duration (wks)}} \right]$$

Figure 2. Calculation of the Relative Dose Intensity (RDI)

After intended regimen is determined, the RDI for each drug received can be calculated, and then averaged across drugs to obtain the ARDI. Participants who discontinue therapy early will only contribute data (for both the numerator and denominator) for the cycles they complete, so as to not obscure the intentional dose reductions we intend to capture by conflating dose intensity with discontinuation. Notably, by including information on observed and planned treatment duration, the RDI (and ARDI) inherently account for treatment delays.

2c3.4. AIM 1. Statistical Analysis. Initial analyses will examine the main effect of the association of BMI with the dosing variable outcomes. In analyses that examine predictors of dose reduction within BMI strata, BMI at diagnosis will be categorized into normal weight (BMI 18.5−<25 kg/m^2), overweight (BMI 25−<30 kg/m^2), obese class I (BMI 30−<35 kg/m^2), obese class II or greater (BMI ≥35 kg/m^2); we will explore whether we can further stratify into obese class III or greater (BMI ≥40 kg/m^2) and retain robust estimates. Interaction between each predictor and BMI will be tested using a likelihood ratio test. Associations will be evaluated overall, as well as by regimen and individual drug.

Both outcome variables (the first cycle dose proportion and the ARDI) will be modeled as continuous linear variables, representing the proportion received relative to that expected. Given the use of a continuous outcome, analyses will be conducted using linear regression. In minimally-adjusted analyses of each individual potential predictor, we will adjust for a core set of covariates, including age, race/ethnicity, and study site (KPNC vs GH). In fully adjusted analyses, we will include all variables in which the p-value<0.10 for associations in the minimally-adjusted models, so as to mutually adjust effect estimates and distinguish variables that are independently associated with dosing.

In secondary analyses, we will alternatively examine associations using a binary outcome, defining dose reductions as receipt of <85% of the expected dose. This cut-point was selected based on research suggesting that patients experience reduced survival with a RDI <85% (31). In order to address concerns regarding the subjectivity involved in determining expected regimen by manual review/chart abstraction, sensitivity analyses will be conducted excluding participants for whom the intended regimen was determined manually. We will explore whether associations vary by race/ethnicity, as African-Americans also are known to have lower neutrophil counts (49), but dose reduction due to neutropenia is not race-specific, and it is even more likely that chemotherapy dose reductions may occur in this group, resulting in lower-than-optimal therapeutic doses being administered. It is therefore possible that the predictors of dose reductions will vary by racial/ethnic group. Lastly, we will conduct sensitivity analyses among women less than age 70 y, given that older women do not uniformly receive chemotherapy and the predictors of chemotherapy dose reductions may vary among older populations.

2c3.5. AIM 1. Statistical Power. Preliminary data indicate that approximately 11,325 cases will have received chemotherapy, and a standard deviation of 0.045 for chemotherapy dosing. If we conservatively assume a larger standard deviation of 0.2 and α=0.05, we have still have excellent

Table 5. Power for Analyses of Predictors of Dose Intensity

Exposure Prevalence	Unexposed (E_-) N	Exposed (E_+) N	E_- Mean Dose Intensity	E_+ Mean Dose Intensity	SD	Power
Overall (n=11,325)						
Q5 vs Q1	2,265	2,265	0.98	0.93	0.2	>0.99
Obese Class I (30-≤35 kg/m^2) (n=2,176)						
Q5 vs Q1	435	435	0.98	0.93	0.2	0.96
Obese Class II (35 kg/m^2) (n=1,709)						
Q5 vs Q1	342	342	0.98	0.93	0.2	0.90

ABBREVIATIONS: SD (standard deviation); Q5 vs Q1 (comparing quintile 5 vs quintile 1); 50% vs 50% (comparing persons above and below the median)

power (**Table 5**). For example,overall, we have more than 99% power to detect an association between exposure (categorized in quintiles)and dose intensity. Within individual BMI groups (such as the obese class I and obese class II+), we have excellent power to detect an association between variables, even when categorized in quintiles, and dosei ntensity.

2c4. AIM 2. In Aim 2, we will evaluate the associations between body size (BMI) at the time of diagnosis and breast cancer recurrence and survival, and the extent to which these associations are mediated by dose reductions in chemotherapy (**Figure 3**).

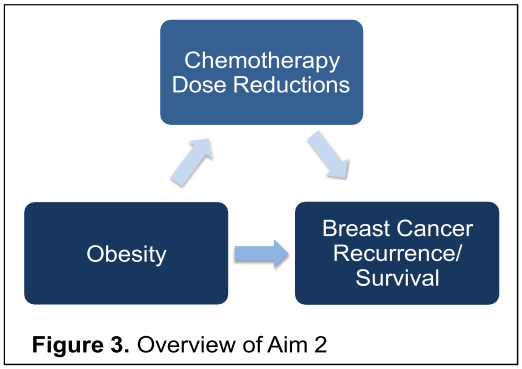

Figure 3. Overview of Aim 2

2c4.1. AIM 2. Study Population. Analyses will include approximately 33,908 women diagnosed with Stage I-IIIA breast cancer.

2c4.2. AIM 2. Exposure. At KPNC, BMI at diagnosis will be captured using weight closest to diagnosis (no more than 6 months prior and no more than 2 months after diagnosis, provided prior to treatment). At GH, BMI at diagnosis will be captured using the weight recorded immediately before the start of treatment. At each site, height will be pulled from the record closest to diagnosis. In primary analyses, BMI (kg/m^2) will be categorized as underweight (<18.5), normal weight (18.5-<25), overweight (25-<30), obese class I (30-<35) and obese class II+ (35+). In secondary analyses, we will disaggregate class II obesity (35-<40) from obese class III+ obesity (40+).

2c4.3 AIM 2. Mediator. Primary analyses will focus on first-cycle dose reductions as the mediator, given that we want to capture mediation by intentional dose reductions (7), without incorporating dose modification due to toxicity and delay. Dose-reduced chemotherapy, as defined by receipt of <85% BSA-determined dose, will be categorized as a three-level variable: fully dosed (receipt of chemotherapy and receipt ≥85% of the expected dose; referent), dose reduced (receipt of chemotherapy and receipt <85% of the expected dose), and did not receive chemotherapy. Note that among those receiving chemotherapy, dose intensity will be treated as a binary variable (rather than a continuous variable), as we are using mediation analyses to estimate what the association between obesity and adverse breast cancer outcomes would have been had obese women not been dose reduced. Furthermore, we have included a category corresponding to no receipt of chemotherapy so that analyses address the question of mediation between obesity and adverse outcomes among the entire population of breast cancer patients, not just those receiving chemotherapy. However, in secondary analyses, we will restrict the population to those receiving chemotherapy and will evaluate this question among those receiving chemotherapy; in this secondary analysis, dose reduction will be defined as a simple binary variable fully dosed (receipt of chemotherapy and receipt ≥85% of the expected dose; referent), dose reduced (receipt of chemotherapy and receipt <85% of the expected dose). In secondary analyses, we will also evaluate mediation by the ARDI, as opposed to first-cycle dose reduction.

2c4.4. AIM 2. Outcome. The outcomes in this analysis will be both survival and recurrence, as detailed below. At both sites, women will be followed for recurrence and survival through 2019. In KPNC, on average, women will have the opportunity to be followed for recurrence and survival for 6.5 years (up to 13 years) and at GH, on average, women will have the opportunity to be followed for 9.5 years (up to 15 years).

2c4.4a. Survival. Participants will be followed for death (any cause of death, as well as breast cancer specific death). At KPNC and GH, deaths are available through the VDW, as identified through sources such as automated linkage to state death tapes and enrollment data. Preliminary data indicate that we might expect 4,282 deaths, including 3,682 deaths at KPNC and 600 deaths at GH. Cause of death, as recorded in the VDW, is determined by death certificates.

2c4.4b. Recurrence. Recurrence will be defined by locoregional recurrences and distant recurrences/ metastases, as identified by algorithms using data from several administrative sources. This approach, developed at GH in Stage I-II breast cancer cases (50), was extended in KPNC to include Stage I-IIIA cases (51). Based on the findings of these papers, we will apply a high-quality high-sensitivity algorithm to identify potential recurrences, after which we will: i) verify all recurrences with medical chart review and ii) verify a random subset of non-recurrences (n=100 at each site) to ensure the algorithm performs well. Although we plan to follow all participants for recurrence through 2019, verified recurrence data are available through 2011 for the original COMBO cohort (~82% of GH participants) and through 2015 for ~20% of KPNC participants (through the Pathways Study). These data can act as the gold standard to ensure that our approach is performing as expected. We expect approximately 3,281 recurrences, including 2,750 recurrences at KPNC and 531 recurrences at GH.

2c4.5. AIM 2. Statistical Analysis. We will use Cox proportional hazards regression to evaluate how BMI relates to recurrence, mortality, and breast cancer-specific mortality, and will use inverse probability weighted

(IPW) marginal structural Cox proportional hazards models to estimate the controlled direct effect of obesity onboth recurrence and survival that is not mediated by dose reductions in chemotherapy (37, 38). This approach, as compared to the traditional approach in which one uses the change in the beta coefficient in models with and without a potential mediator to assess mediation (52), provides several methodologic advantages. Most importantly, the marginal structural model approach allows for adjustment for endogenous confounding (53)—meaning that our exposure, obesity, may affect variables such as comorbidity, which can act to confound the mediator-outcome (dosing-survival) relationship. In the course of building the marginal structural models, we will estimate two sets of stabilized IPW using multinomial logistic regression. The first corresponds to the exposure (to account for measured confounding of the exposure-outcome relation). The second corresponds to the mediator (to account for measured confounding of the mediator-outcome relation). After deriving these weights, they will be combined to obtain an overall weight, which will then applied to the proportional hazards regression model to adjust for confounding. When building weights to account for confounding, we will adjust for variables such as: study site, age, race/ ethnicity, area-based socioeconomic measures, menopausal status, comorbidities (diabetes, cardiovascular disease, renal impairment, hepatic impairment), calendar time, smoking status, and cancer characteristics (stage, grade, nodal status, ER status).This list of covariates will be modified based on the results of Aim 1 to ensure that we include all identified factors associated with dosing and which are known to be associated with survival. In analyses of mortality, participants will be followed until death, disenrollment (defined as an enrollment gap exceeding 90 days), or end of study (2019), whichever occurs earliest. In analyses of recurrence, participants will be followed until recurrence, disenrollment (defined as an enrollment gap >90 days), death, or end of study (2019), whichever occurs earliest.

In secondary analyses, we will: i) further disaggregate those with a BMI 35+ kg/m^2 into those who are class II obese (BMI: 35-<40 kg/m^2) and class III or greater obese (BMI: 40+ kg/m^2) to examine the extent to which dose reductions mediate this association at different levels of obesity, ii) evaluate mediation specifically among the subset of individuals receiving chemotherapy, and iii) evaluate mediation by the ARDI. We will also evaluate whether the association varies by select variables for which we hypothesize that the mediated effect may vary by strata, including: comorbidity (such as diabetes or cardiovascular disease), age, race/ethnicity cancer stage, ER status, and calendar time. For example, we might expect that the mediation by dose reductions may have been stronger in earlier years, when there was less clinician awareness of this issue. Similarly, we might expect stronger mediation among ER- patients, given a prior study that that dose reductions of CMF impacted survival specifically among the ER- patients, but not among ER+ patients (31). We will also explore associations by race/ethnicity, given the known disparities in breast cancer outcomes with poorer survival among African-American women, along with the greater prevalence of obesity, neutropenia, and dose reduction in this group (19, 49, 54, 55).

Sensitivity analyses will be conducted restricting to those diagnosed on or before 2014 to ensure that the association is comparable when restricted to women who would have the opportunity to be followed for ≥5 years. In 2013, KPNC began routinely assessing physical activity and alcohol consumption. In sensitivity analyses restricted to those with this information available, we will further adjust for both physical activity and alcohol consumption to evaluate whether inclusion of these variables meaningfully affects results. Lastly, we will also explore mediation by different thresholds for dose reductions (<90%; <70%), as this may ultimately provide insight into the level of dose reduction that may be tolerated before compromising breast cancer outcomes.

2c4.6. AIM 2. Statistical Power. As can be seen in **Table 6**, we are well powered for analyses of both recurrence and survival. To conservatively

Table 6. Power Calculations for Aim 2

Outcome	Total N	Exposed N	Unexposed N	Comparison N	Events N	HR	Power (Inflation Factors of 0.7, 0.8, 0.9)*
Obese Class I (BMI 30-≤35 kg/m^2) vs Normal Weight (BMI 18.5-≤25 kg/m^2)							
Survival	33,908	6,519	10,990	17.509	2,211	1.4	>0.99 (>0.99, >0.99, >0.99)
Recurrence	33,908	6,519	10.990	17,509	1,694	1.4	>0.99 (>0.99, >0.99, >0.99)
Obese Class II+ (BMI 35+ kg/m^2) vs Normal Weight (BMI 18.5-≤25 kg/m^2)							
Survival	33,908	5,093	10,990	16,083	2,031	1.4	>0.99 (>0.99, >0.99, >0.99)
Recurrence	33,908	5,093	10.990	16,083	1,556	1.4	>0.99 (>0.99, >0.99, >0.99)

*Denotes power when effectively reducing the sample size (and cases) to 70%, 80%, and 90%

account for the loss of power when conducting mediation analyses, we have applied inflation factors ranging from 0.7-0.9 (representing 70% and 90% of the total sample size and number of cases). Assuming an overall hazard ratio (HR) of 1.4 when comparing obese class I individuals to normal weight individuals, we have excellent power for mediation (1); the same holds for analyses of obese class II+ individuals. Given the very large number of events, power analyses further indicate that we will have adequate power for stratification by variables even with relatively unbalanced strata; for example, if we assume that ER- cases account for approximately 16% of cases, and conservatively assume an inflation factor of 0.8 for mediation analyses, we will still have 81% power for analyses evaluating the extent to which the association between class I obesity and survival is mediated by dose reduced chemotherapy among ER- cases.

2c5. AIM 3. In Aim 3, we will address the question of whether the risk of toxicity associated with full BSA-determined dosing varies across the spectrum of BMI.

2c5.1. AIM 3. Study Population. This analysis will be focused on the estimated 11,155 women receiving adjuvant, first-line chemotherapy with a BMI ≥18.5 kg/m^2. Underweight women (BMI <18.5) will be excluded due to small sample size and potential heterogeneity with the normal weight group.

2c5.2. AIM 3. Exposure. We are specifically interested in the relationship between BMI and toxicity among women receiving the full BSA-determined dose of chemotherapy so that we can better understand whether fully-dosed class I obese women (BMI 30-<35) and fully-dosed class II+ obese women (BMI 35+) experience excess toxicity as compared to their fully-dosed non-obese counterparts (BMI 18.5-<30).

In these analyses, determination of whether a woman has received her full BSA-determined dose will be defined in two ways: first-cycle dose proportion and the ARDI. A woman undergoing chemotherapy has the potential to transition to and from being 'fully-dosed' as she moves through each cycle of treatment. This has two methodological consequences for analyses in which the full BSA-determined dose is determined by the ARDI (a measure updated at each treatment cycle). First, although we are specifically interested in the comparison of fully-dosed obese women and fully-dosed non-obese women, rather than simply restricting the analyses to fully-dosed women, we will include all women in the analysis regardless of dosing, to allow these individuals to transition to and from the fully-dosed group as she moves through cycles of chemotherapy. Second, it is critical that we use a time-varying variable to capture changes in dosing at each cycle. This will ensure that analyses only incorporate information on dosing before the development of toxicity, preserving the temporal relationship between the exposure and outcome. Therefore, exposure will be defined as follows: fully-dosed non-obese (referent; ARDI ≥85%; BMI 18.5-<30), dose-reduced non-obese (ARDI <85%; BMI 18.5-<30), fully-dosed obese class I (referent; ARDI ≥85%; BMI 30-<35), dose-reduced obese class I (ARDI <85%; BMI 30-<35), fully-dosed obese class II+ (ARDI ≥85%; BMI 35+), and dose-reduced obese class II (ARDI <85%; BMI 35+). In analyses of the ARDI, this exposure will be updated with each cycle of chemotherapy (reflecting cumulative dose intensity).

2c5.3. AIM 3. Outcome. The outcome in this analysis is a composite measure of severe toxicity, including neutropenia, neuropathy, renal impairment, hepatic impairment, and cardiotoxicity, defined as Grade 3+ toxicity, according to the *NCI Common Terminology Criteria for Adverse Events* (NCI CTCAE) version 4.0. Data will be pulled from a combination of ICD-9/10 diagnosis codes, laboratory values, and drug codes (20). For example, neutropenia will be defined using relevant diagnosis codes, laboratory values for neutrophil count, and use of granulocyte colony stimulating factor (G-CSF), as defined in a prior KPNC study (20).

2c5.4. AIM 3. Statistical Analysis. We will use Cox proportional hazards regression to evaluate the association between BMI and toxicity, with a particular interest in the comparison of fully-dosed obese women to the fully-dosed non-obese women. Participants will be followed from the date of first chemotherapy until development of toxicity, disenrollment (defined as an enrollment gap exceeding 90 days), recurrence, diagnosis with a new cancer, death, or end of study (2019). In analyses in which dosing is defined by first cycle dose reductions, exposure will be modeled at baseline, and in analyses in which dosing is defined as a time-varying variable, the BMI-ARDI category will be updated with each cycle of chemotherapy. Analyses will be adjusted for the following covariates: study site, age, race/ethnicity, area-based socioeconomic measures, pre-existing comorbidities (diabetes, heart disease, renal impairment, hepatic impairment), treatment, calendar time, stage, and hormone receptor status.

Secondary analyses will examine associations for subsets of toxicity (e.g., hematologic vs non-hematologic toxicity), as well as common individual toxicities (e.g., neutropenia, neuropathy). We will also explore whether associations vary by treatment (e.g., dose-dense therapy vs not dose-dense therapy), age, race/ethnicity, and comorbidity. We will further explore the associations with overall toxicity when disaggregating class II (BMI 35-≤40 kg/m^2) and class III obesity (BMI 40+ kg/m^2).

In sensitivity analyses, we will: i) disaggregate the non-obese reference group into the normal weight (18.5-

<25) and overweight (25-<30) to confirm the appropriateness of collapsing these groups into a single reference category, ii) evaluate if the association varies when applying different thresholds to define fully-dosed chemotherapy (e.g., ≥90%, ≥95%), iii) will evaluate if the association changes when censoring all participants 6 months after the completion of chemotherapy (to address whether the association is comparable for short-term toxicity), and iv) in time-varying analyses, we will evaluate if the associations vary when using a simple updated cycle-specific dose proportion (to capturing acute effects on toxicity), rather than cumulative dose intensity, as captured by the ARDI.

2c5.5. AIM 3. Statistical Power. We have excellent power when comparing fully-dosed obese class I women with fully-dosed normal-weight women, across a range of expected outcomes (**Table 7**). For example, if 50% of non-obese fully-dosed women develop toxicity over the course of follow-up (56), we are powered to detect a HR as small as 1.11. We are also well powered to detect individual outcomes or subsets of outcomes; for example, when comparing fully dosed obese class I women to fully-dosed non-obese women, we can detect a HR of 1.22 for outcomes that occur in 15% of the study population. Power is relatively similar for class II+ obesity. In exploratory analyses in which we will further disaggregate class II (BMI 35-<40 kg/m^2) and class III+ obesity (BMI≥ 40 kg/m^2), we are powered to detect an association of 1.21 for any toxicity (assuming this occurs in 50% of the population and assuming class III obese persons to comprise ~6% of the total study population), and are powered to detect a HR of 1.33 for toxicities occurring in 25% of the population.

Table 7. Power for Toxicity Analyses, for Both Overall Obesity and Severe Obesity

Non Obese Fully Dosed N	Probability of Toxicity among Normal Weight	Obese Class I (BMI 30-<35) vs Non-Obese (18.5-<30)				Obese Class II+ (BMI 35+) vs Non-Obese (18.5-<30)			
		Obese I Fully Dosed N	Total Fully Dosed N	Fully Dosed Event N	HR Powered to Detect	Obese II Fully Dosed N	Total Fully Dosed N	Fully Dosed Event N	HR Powered to Detect
7,047	50%	1,932	8.979	4,421	1.11	1,367	8,414	4,151	1.13
7,047	25%	1,932	8.979	2,186	1.17	1,367	8,414	2,055	1.20
7,047	15%	1,932	8.979	1,298	1.22	1,367	8,414	1,223	1.27
7,047	10%	1,932	8.979	858	1.29	1,367	8,414	809	1.35
7,047	5%	1,932	8.979	420	1.45	1,367	8,414	397	1.57

2c6. POTENTIAL CHALLENGES & STRATEGIES TO ADDRESS THESE CHALLENGES. Our study populations cover a wide span of time (26 years), with GH cases diagnosed between 1994-2015 and KPNC cases diagnosed from 2006-2019, posing a concern that there could be substantive differences between groups. However, preliminary data indicate that distributions of key variables are relatively comparable across sites; even so, we will adjust for study site in all analyses. Another consideration in this observational study is that patients will be receiving heterogeneous chemotherapy regimens; however, this is representative of treatment as received in the real-world setting. Where applicable, we will examine associations by individual regimens and drugs. Given that we are using data from integrated healthcare delivery settings, we plan to use diagnostic codes, laboratory results, and prescription data to define toxicities, and it is possible that we may miss toxicities, especially lower grade toxicities, that are not captured by these sources. In the initial COMBO study, toxicities are also captured by medical record abstraction and therefore within this cohort, we can assess the agreement between the two approaches. Similarly, given the data source, we do not universally have data on all potential covariates. However, given that KPNC began assessing alcohol consumption and physical activity in 2013, we will be able to conduct sensitivity analyses assessing adjustment for physical activity and alcohol consumption in the subset of KPNC participants diagnosed after 2013 (for whom this information is available in the VDW). Furthermore, in the smaller subset of KPNC participants (n=4,505) who were part of the Pathways Study, a KPNC-embedded study that also included rich questionnaire data, we will conduct sensitivity analyses to assess the impact of additional adjustment for a whole range of factors. Lastly, when evaluating the interrelationship between dosing and obesity as they relate to toxicities, we will not be powered to interpret comparisons of dose-reduced groups to fully-dosed groups within each BMI category, given the relatively small number of individuals who are dose reduced within each BMI category. However, we are very well powered to address our question of interest – whether fully dosed obese women experience excess toxicity compared to fully-dosed normal weight women.

2c6. STUDY TEAM. This study will be conducted by a multi-site and interdisciplinary team.

Elizabeth Kantor, Principal Investigator of the proposed study, is a cancer epidemiologist in the Department of Epidemiology and Biostatistics at Memorial Sloan Kettering Cancer Center (MSKCC). Elizabeth is part of the Cancer Research Network (CRN) Scholars Program, a 26-month mentored program to help early-stage investigators become familiar with conducting research in integrated healthcare delivery systems and to launch a large research project within the CRN.

This project will be pursued in collaboration with a team of experienced investigators at KPNC, GH, and Rutgers/Cancer Institute of New Jersey, with further input provided by consultants from the University of Michigan and the Harvard T.H. Chan School of Public Health. Together, this team will provide expertise on all aspects of the project, from use of data in an integrated healthcare delivery system, to topical guidance on obesity, pharmacy, medical oncology, chemotherapy dosing, toxicity, and breast cancer epidemiology. Notably, several investigators involved with the proposed study were involved in our preliminary project (Drs. Kantor, Kushi, and Bandera); this experience in extracting and analyzing these data will be of considerable value given the complexity and richness of these data. **Dr. Larry Kushi** is an epidemiologist at KPNC and is the PI of the CRN (U24 CA171524) and the Pathways Study (R01 CA105274, U01 CA195565) in which pilot studies were conducted and which will serve as a validation subcohort. He has published extensively on breast cancer using KPNC data as well as prior to joining KPNC 15 years ago, and has recently published on chemotherapy dosing of ovarian cancer patients in KPNC in collaboration with co-investigator, Dr. Elisa Bandera (20, 48). **Dr.Tatjana Kolevska** is the Chief of Medical Oncology for the Napa-Solana region of KPNC, and also the Chair ofthe Chiefs of oncology for all of KPNC. She is also co-lead of the KP Oncology Clinical Trials group. She will provide clinical input on questions related to medical oncology and dosing practices at KPNC. **Dr. Candyce Kroenke**, an epidemiologist at KPNC who led a recent recurrence algorithm study (51), will offer input throughout the course of the project and will support the application of the algorithm in identification of recurrences. **Erin Aiello Bowles**, a breast cancer epidemiologist at GH, has led a number of breast cancer epidemiology studies, and several multi-site studies in the CRN on chemotherapy and cardiotoxicity (44, 57-59). **Dr. Denise Boudreau**, a pharmacoepidemiologist and breast cancer epidemiologist at GH and PI of the COMBO study (R01 CA120562) will provide expertise with regard to the COMBO data, as well as pharmacy questions that arise. **Dr. Elisa Bandera** is a physician and Professor of Epidemiology and Co-Leader of Cancer Prevention and Control at Rutgers Cancer Institute of New Jersey. Dr. Bandera is leading a cohort of breast cancer survivors evaluating the impact of obesity-related comorbidities and survival, and has recently led projects on obesity and race/ethnicity in relation to dosing of chemotherapy among KPNC ovarian cancer patients in collaboration with Dr. Kushi (20, 48). Further support will be provided by consultants, Dr. Jennifer Griggs and Dr. Eric Tchetgen Tchetgen. **Dr. Jennifer Griggs** is a medical oncologist with an extensive history in the field of chemotherapy dosing. She conducted early work showing that obese women are more likely to be dose-reduced than normal weight women, and she was the lead author of the 2012 ASCO guidelines. She will therefore be an ideal resource for the proposed project. **Dr. Eric Tchetgen Tchetgen** is a Professor of Biostatistics and Epidemiologic Methods at the Harvard T.H. Chan School of Public Health. He has published extensively on causal mediation methods, and will therefore be able to advise as we develop, implement, and interpret the results of our mediation models.

2c8. STRENGTHS & SUMMARY. This study has several important strengths. Most importantly, we are leveraging our large sample size and rich data to address unanswered questions that have the potential to affect clinical care. Specifically, through use of integrated healthcare delivery systems, we are able to use rich treatment data, including information on the specific drugs given, doses administered, and dates on which treatment was received. Furthermore, we are able to obtain data on comorbidities, toxicities, and long-term health outcomes, including recurrence and survival. These are critical variables to this project, and cannot be obtained in cancer registries or most studies of survivorship. As this project will be conducted within two integrated healthcare delivery systems, patients have comparable access to care, reducing concern about residual confounding by access to care. This study will shed light on optimal chemotherapy dosing practices among obese women, and has the potential to affect clinical practice and improve health outcomes for the 102,000 obese women diagnosed with breast cancer each year in the US.

2c9. TIMELINE

Activity	Lead Site	Yr 1		Yr 2		Yr 3		Yr 4		Yr 5	
Coordinate All Activities	MSKCC										
Obtain IRB Approvals and Execute Data Use Agreements	All										
Abstract BMI and chemotherapy data on all GH participants	GH										
Abstract electronic chemotherapy data	KPNC										
Bin participants into intended chemotherapy regimens	MSKCC										
Medical chart abstraction chemotherapy data	KPNC										
Abstract remaining data	GH/KPNC										
Develop/validate recurrence algorithms	GH/KPNC										
Analyze Data/Draft Paper: Aim 1	MSKCC										

Analyze Data/Draft Paper: Aim 2	MSKCC										
Analyze Data/Draft Paper: Aim 3	MSKCC										

2c9. TIMELINE (Alternative Format)

Activity	**Lead Site**	**Yr 1**		**Yr 2**		**Yr 3**		**Yr 4**		**Yr 5**	
Coordinate All Activities	MSKCC	X	X	X	X	X	X	X	X	X	X
Obtain IRB Approvals and Execute Data Use Agreements	All	X									
Abstract BMI and chemotherapy data on all GH participants	GH		X	X							
Abstract electronic chemotherapy data	KPNC		X	X							
Bin participants into intended chemotherapy regimens	MSKCC			X	X	X					
Medical chart abstraction chemotherapy data	KPNC					X	X				
Abstract remaining data	GH/KPNC		X	X	X	X	X				
Develop/validate recurrence algorithms	GH/KPNC					X	X	X	X		
Analyze Data/Draft Paper: Aim 1	MSKCC							X	X		
Analyze Data/Draft Paper: Aim 2	MSKCC								X	X	X
Analyze Data/Draft Paper: Aim 3	MSKCC								X	X	X

PROTECTION OF HUMAN SUBJECTS

This study will use data from two integrated healthcare delivery systems: Kaiser Permanente Northern California (KPNC) and Group Health (GH). As of February 1, 2017, GH became part of Kaiser Permanente, and is now known as Kaiser Permanente of Washington; in this application, we continue to refer to this healthcare system as Group Health. KPNC and GH are members of the Cancer Research Network (CRN), a consortium of research centers affiliated with integrated health care delivery systems. At all CRN sites, electronic data are extracted from clinical and administrative databases and stored for research purposes in the Virtual Data Warehouse (VDW), a standardized data model with common database structure, variable names, and definitions. In this observational study, much of the data will be obtained from the VDW, and data will also be abstracted from medical records. This observational study does not involve an intervention, nor does it involve any patient contact.

The sections below describe risks to human subjects, adequacy of protection for human subjects, potential benefits of the proposed research, and the importance of the knowledge to be gained.

RISKS TO HUMAN SUBJECTS

3a. Human Subjects Involvement and Population Characteristics

Our study population will be comprised of an estimated 33,908 women, ages 18+ y, diagnosed with Stage I-IIIA breast cancer, about 28,546 of whom are members of KPNC and 5,362 of whom are members of GH.

KPNC is one of the oldest and largest integrated healthcare delivery systems in the United States. KPNC provides care to approximately 3.9 million members throughout the San Francisco Bay Area and the Central Valley of California, across a network of 21 hospitals and over 200 outpatient clinics. GH is an integrated healthcare delivery system providing care to ~620,000 members in Washington State and parts of Idaho. A large portion of GH's service area is located within the Western Washington Cancer Surveillance System region, a member of the SEER Program.

We will include women diagnosed with Stage I-IIIA breast cancer diagnosed between 2006 and 2019 at KPNC and between 1994 and 2015 at GH. Participants will be included without regard to race or ethnicity. We will only access data pertaining to patients relevant to this study.

3b. Sources of Materials

At each site, automated data will be collected and data will be further abstracted from medical records. As noted above, we will use electronic data stored in the VDW, a standardized data model with data derived from clinical and administrative databases. The standardized data model provides consistent variable definitions and formats across multiple sources within each site (KPNC and GH), and facilitates harmonization across sites. Each site maintains its own data, yet data are easily shared across sites using standard naming and coding. Data are stored in domains, including neighborhood/contextual data from the Census, demographics, enrollment, encounters, providers, diagnoses, procedures, laboratory results, pharmacy, infusion medications, tumor registry, vital signs, social history, and death. At both sites, we will also collect data from medical records. For example, at GH, we will need to abstract chemotherapy dosing data and height/weight before the start of chemotherapy from all breast cancer cases and at both sites, we will need to abstract data to validate recurrences.

Data collection at KPNC will be overseen by site PI, Larry Kushi, and data collection at GH will be overseen by site PI, Erin Aiello Bowles.

3c. Potential Risks

The main risk to the subjects is a breach of confidentiality. However, we will follow standard practices to protect confidentiality, as described in the sections below.

ADEQUACY OF PROTECTION AGAINST RISKS

Recruitment and Informed Consent

Subjects fitting study inclusion criteria will be identified by programmers at each site; these programmers will pull automated data from the VDW on participants, which will be supplemented with medical record abstraction as needed. This study involves no subject contact, and at KPNC and GH, informed consent is typically not required for use of automated data and chart abstraction for research purposes with prior approval by the Institutional Review Board (IRB) and if the data are not used in a manner in which individuals are identifiable. In accordance with state and federal regulations, we will request a waiver of consent from the IRBs of KPNC and GH to access information from electronic data files and patients' medical records.

Protection Against Risk

This study will involve no direct contact with human subjects, and data used in this project are solely for research. There are no medical risks involved for subjects, and inclusion in this study will not alter participants'

delivery of care. Therefore, this study involves a low level of risk. As noted previously, the main risk is a breach in confidentiality. We will protect against this risk in a number of ways.

First, all study activities will be subject to approval from the IRBs of KPNC and GH (where the data will be collected), as well as MSKCC (where the data will be analyzed); no study activities will commence until approved by the IRBs at these sites. IRB approvals will be based on specific project descriptions; any changes to the original project description will require re-review and re-approval by the IRB.

Secondly, analytic datasets will be only accessible to the study team. We will we not release personal patient information; preliminary or final results will not be released or published with identifying information. Results will not be reported on an individual level: all epidemiologic data will be presented as statistical summaries, and no statistics will be published based on fewer than 5 subjects.

Thirdly, at KPNC and GH, identifiable and protected health information is stored in secured areas with restricted access. At both KPNC and GH, computerized data are protected from unauthorized access by their respective data security policies. Access to information stored on computers is limited to individuals who have been granted rights, and data are protected from unauthorized access by anyone outside by a firewall and virus blocking software. Firewall-protected servers are located in a locked room in a secured facility.

Furthermore, confidentiality of participants will be protected by only sharing a HIPAA-defined limited data set with MSKCC for analysis. Given the use of a limited dataset data, MSKCC investigators will not be able identify or link back to the identities of study participants. Specifically, analytic datasets will exclude direct patient identifiers such as name or medical record number. Only study personnel at each site who need to link data sources, abstract data, or track participants for follow-up purposes will have access to the identifying information for individual study participants. Furthermore, only the staff abstracting the medical records will have access to the medical records, though there will be some allowance for adjudication (if necessary) by the project manager and investigators at the respective sites.

When abstracting data from automated files and medical records, programmers will link records using members' unique medical records number. Study subjects will then be assigned a unique study identification number, and the medical record number will be removed from analytic dataset (which will be stored as a password-protected file). The analytic dataset will not include identifiers such as medical record number, name, or address. However, it will not be strictly be de-identified, as it will include dates of service (e.g., dates of diagnosis, dates of administration of chemotherapy). The analytic dataset will include study-specific identification numbers in case modifications to the analytic dataset are necessary. However, linkage of the study identification number with medical record number will be stored separately and kept at KPNC or GH. Only study personnel at these study sites will have access to their respective linkage files, and these files will not be shared outside these institutions. No personally identifiable information (protected health information) except the minimum necessary for study purposes will be shared with investigators or staff outside KPNC and GH. At all sites, data will be password-protected in subdirectories (for which there is controlled access).

At each site, data handling procedures are clearly documented and all new employees receive training on data handling procedures, confidentiality, and security. After data are abstracted at each site, a limited dataset will be shared with MSKCC for analysis via a secured portal which is encrypted per HIPAA and HITECH safe harbor standards. Only persons involved with this study will have access to files that are transferred using the secure file transfer site.

Lastly, all staff at KPNC and GH have successfully completed HIPAA training as required by KPNC and GH and sign a contract at the time of hire in which they promise to keep confidential all materials with which they come in contact. All personnel who will handle the human subjects data have undergone human subjects training.

POTENTIAL BENEFITS OF THE PROPOSED RESEARCH TO HUMAN SUBJECTS AND OTHERS

Although subjects will not benefit directly from this study, their participation can help inform clinicians treating breast cancer patients. The risks to study subjects will be minimal in relation to the potential benefits to society, and as noted above, many safeguards will be put into place to maintain patient confidentiality.

IMPORTANCE OF THE KNOWLEDGE TO BE GAINED

Obese breast cancer patients are more likely than normal weight patients to receive less than their prescribed dose of chemotherapy due clinician concern about inducing chemotherapy-associated toxicity. Despite recommendations against dose reductions, evidence suggests that this practice continues to occur in obese women. In this project, we propose to i) examine the relationship between body size and chemotherapy dose intensity, and will further examine how the factors contributing to dose reductions vary by body size, ii) evaluate whether dose reductions in chemotherapy mediate the associations between obesity and adverse

breast cancer outcomes, and iii) among women receiving the full dose of chemotherapy, we will evaluate the relationship between body size and toxicity. Addressing these unanswered questions will provide the evidence needed to better inform clinicians and improve clinical care for the estimated 102,000 obese women diagnosed with breast cancer each year in the United States.

7. Data Safety and Monitoring Plan

As this is not an intervention study, a Data Safety and Monitoring Plan is not required.

8. ClinicalTrials.gov Requirements

As this is not an intervention study, registration in ClinicalTrials.gov does not apply.

INCLUSION OF WOMEN AND MINORITIES

As breast cancer is rare in men, this study will be focused on female breast cancer patients. Participants will be selected without regard to race/ethnicity. Overall, we expect 67.8% of participants to be non-Hispanic white (65.0% at KPNC and 83.0% at GH), 7.0% to be non-Hispanic black (7.8% at KPNC and 2.9% at GH), 13.2% to be Asian (14.7% at KPNC and 4.8% at GH), and 10.8 % to be Hispanic (11.7% at KPNC and 6.1% at GH).

PHS INCLUSION ENROLLMENT REPORT

This report format should NOT be used for collecting data from study participants.

OMB Number:0925-0001 and 0925-0002
Expiration Date: 10/31/2018

***Study Title:** Obesity, Chemotherapy Dosing, and Breast Cancer Outcomes

***Delayed Onset Study?** ❑ Yes ☑ No

If study is not delayed onset, the following selections are required:

Enrollment Type ☑ Planned ❑ Cumulative (Actual)

Using an Existing Dataset or Resource ☑ Yes ❑ No

Enrollment Location ☑ Domestic ❑ Foreign

Clinical Trial ❑ Yes ☑ No

NIH-Defined Phase III Clinical Trial ❑ Yes ☑ No

Comments:

Racial Categories	Ethnic Categories									Total
	Not Hispanic of Latino			Hispanic or Latino			Unknown/Not Reported Ethnicity			
	Female	Male	Unknown/ Not Reported	Female	Male	Unknown/ Not Reported	Female	Male	Unknown/ Not Reported	
American Indian/Alaska Native	233	0		28	0					261
Asian	4462	0		193	0					4655
Native Hawaiian or Other Pacific Islander	165	0		10	0					175
Black or African American	2369	0		55	0					2424
White	23024	0		3369	0					26393
More than One Race	0	0		0	0					0
Unknown or Not Reported										
Total	30253	0		3655	0					33908

INCLUSION OF CHILDREN

As this is a study of breast cancer and breast cancer is extremely rare in children, this study will be limited to adult breast cancer patients 18 years of age and older.

AUTHENTICATION OF KEY BIOLOGICAL, AND/OR CHEMICAL RESOURCES

No biological or chemical resources, including cell lines, chemicals or any other biological materials, will be involved in the proposed activities.