SF 424 R&R and PHS-398 Specific
Table of Contents

PI: GOMEZ, SCARLETT L	Title: Lung cancer in never smokers: incidence, risk factors, and molecular characteristics in Asian American, Native Hawaiian and Pacific Islander females	
	FOA: PA 13-302	
	FOA Title: RESEARCH PROJECT GRANT (PARENT R01)	
	Organization: CANCER PREVENTION INSTIT OF CALIFORNIA	
Senior/Key Personnel:	*Organization:*	*Role Category:*
Scarlett Gomez Ph.D.	Cancer Prevention Institute of California	PD/PI

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 NO, is the IRB review pending? ☒ Yes ☐ No

2. Are Vertebrate Animals Used? ☐ Yes ☒ No

3. Is proprietary/privileged information included in the application? ☐ Yes ☒ No

4a. Does this project have an actual or potential impact on the environment? ☐ Yes ☒ No

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

6. Does this project involve activities outside of the United States or partnerships with international collaborators? ☐ Yes ☒ No

PROJECT SUMMARY/ABSTRACT

For females of Asian Americans, Native Hawaiians and Pacific Islanders (AANHPI) ethnic groups, lung cancer is the 3rd or 4th most common cancer, but the most common cause of cancer death. The burden of lung cancer among AANHPI females is striking considering their low prevalence of smoking, that more than half of lung cancers occur in never smokers, and contributing risk factors beyond smoking remain largely unknown. Furthermore, the incidence rates of lung cancer, especially adenocarcinoma, are either stable or increasing among Filipina, Korean, and Chinese American females, in stark contrast to the overall declining incidence rates of lung cancer in other U.S. racial/ethnic groups. Given the lack of information on smoking status for populations at risk, population-level incidence rates stratified by smoking status, race/ethnicity, and sex are not available, constituting a critical gap in knowledge. At present, there is no single sufficiently large data source to document lung cancer incidence rates by smoking status among specific AANHPI ethnic groups, which is central to understanding and reducing the burden of disease in this heterogeneous population that includes individuals from more than 30 different countries, speaking more than 100 languages. To address these gaps, the objective of our study is to leverage prospective data from two large electronic health record (EHR) databases, comprising 3.2 million individuals, 1.8 million females, and over 240,000 AANHPI females, with up to 15 years follow-up, to estimate their lung cancer incidence and characterize the epidemiology of lung cancer by specific single and mixed race/ethnicity and smoking status. For Aim 1, we will calculate overall and histological cell-type specific incidence rates of lung cancer by smoking status, and compare the distribution of sociodemographic, tumor (e.g. stage, histology), and molecular characteristics (e.g. EGFR, ALK) of lung cancer cases by race/ethnicity and smoking status. Males and other races/ethnicities will be examined for comparison in Aim 1. For Aim 2, we will conduct a longitudinal analysis of lung cancer risk, including absolute risk modeling, examining six exposure domains: second-hand smoke, previous lung diseases, infections, reproductive history and hormone exposure, body size, and neighborhood environmental factors, including measures of particulate matter, traffic density, neighborhood socioeconomic status, and ethnic enclave. This study will use EHR data from the Northern California Sutter Health system and from Kaiser Permanente Hawaii – each specifically selected for their robust AANHPI representation and high quality data. Coupled with a focus on distinct subpopulations defined by race/ethnicity (including well-defined mixed race/ethnic groups), smoking status, environmental characteristics, and tumor-based molecular markers, this highly efficient study will provide much-needed information on lung cancer risk among AANHPI never smokers, serving as a critical evidence base to inform screening, research, and public health priorities in this growing population.

PROJECT NARRATIVE

This application will quantify the burden of lung cancer among Asian American, Native Hawaiian and Pacific Islanders (AANHPIs) and will identify etiologic risk factors among never-smoking AANHPI women, using information from electronic health records linked with data on neighborhood contextual factors, outdoor air pollution, and from cancer registries. Efficiently leveraging and combining existing data, this project will provide much-needed information for better understanding the epidemiology of lung cancer among AANHPI women, among whom the factors responsible for the cause of this common cancer are largely unknown.

CPIC FACILITIES AND OTHER RESOURCES

ENVIRONMENT

Environment – Contribution to Success:

The Cancer Prevention Institute of California develops and implements innovative cancer prevention research and outreach programs to deliver a comprehensive arsenal for defeating cancer. Established in 1974 as the Northern California Cancer Program, then named the Northern California Cancer Center (NCCC), the organization changed its name in 2010 when it became the Cancer Prevention Institute of California (CPIC), which reflects the organization's broader scope and demonstrates its large scale impact of preventing cancer before it starts. The study team at CPIC has full access to the human and informatics resources necessary to successfully lead this study; these resources are described below. Briefly, these include the Greater Bay Area Cancer Registry and its associated staff, with expertise in managing and analyzing cancer incidence and related data, as well as appropriate computer hardware and software equipment and associated technical support from the organizational Information Technology staff. Through its formal affiliation with the Stanford Cancer Institute, and NCI-designated cancer center, CPIC scientists also have access to other important resources such as the Stanford Cancer Institute cores, including biostatistics support, and other Stanford resources including the full complement of scientific journals and books from the medical and other libraries. The physical location of the CPIC, in the heart of the San Francisco Bay Area, also conveniently places it within easy access to other research organizations, including Stanford University, and Palo Alto Medical Foundation (PAMF). The offices themselves are easily accessible via local and air transportation and can easily be used for hosting meetings. Finally, members of the CPIC team has previously collaborated successfully with Stanford University and PAMF investigators, thus demonstrating their ability to successfully communicate and collaborate on the proposed study. All of these aspects of the CPIC environment and its resources make CPIC an ideal institution for this current application.

FACILITIES:

Information Technology:

The organization's resources consist of a multi-platform communications infrastructure based on:

- Microsoft Windows workstations
  - Centrally managed and controlled security updates
- Microsoft Active Directory for consolidated user account and permissions control
- Microsoft Exchange and Microsoft Office 365 email system
- Cisco networking and security systems:
  - Firewall / Network access system
  - Network core switching
  - Virtual Private Network (VPN) remote connectivity solution
  - Voice over Internet Protocol (VOIP) telephony solution
- Symantec/Veritas backup and restore systems
- Symantec Cloud anti-virus
  - Updated regularly
  - Centrally managed

In addition to the communications infrastructure, appropriate foundational systems and support services are provided for all activities required by this study including:

- Access to ESRI ArcGIS products through Stanford
- Development services (e.g., Visual Basic, Hyper Text Markup Language [HTML], Extensible Markup Language [XML], SAS)
- SAS data analysis and working platform
- Print services, including large high volume printers available to all staff
- Teleconference and videoconference capabilities

CPIC has more than three decades of experience in implementing and securely managing complex data for research purposes. Data at CPIC are stored on secure, firewalled, password-protected servers with limited access to designated study personnel.

Office:

Fremont Office: The organization occupies 18,468 square feet on two floors at the Fremont, CA, site. The space includes numerous furnished private offices and cubicles, four conference rooms, a large training room with a moveable wall that enables it to be transformed into a large open meeting space. Administration, IS/IT, research programs, the Surveillance, Epidemiology, and End Results program, the regional office of the California Cancer Registry program, legal and regulatory affairs, and other education/information programs share this space. Two large-volume photocopiers are available at the Fremont site for rapid copying and collating. Four smaller photocopy machines are also available. Six facsimile machines are also available. The Fremont office is a short drive (<18 miles) across the San Francisco Bay from Stanford University and the PAMF Research Institute. CPIC scientists also have access to office and conference spaces at Stanford. Drs. Gomez, Cheng, Shariff-Marco, DeRouen and Ms. Allen, Yang, and Li have offices at the Fremont location.

Berkeley Office: The organization occupies 3,328 square feet in Berkeley, CA, which is the site of the GIS lab. This space includes several offices, a small computer room and kitchen, and two conference rooms. IS/IT resources include Windows-based, portable and non-portable PC microcomputers with mid- to large-size data storage, management and statistical analyses capabilities, as well as record linkage, mapping, scanning, graphics, desktop publishing, Internet accessibility, and electronic-mail capabilities. The office's PCs are connected to a wide-area network facilitating the communication, research, and data-exchange capabilities among the staff at the Berkeley office, and externally with the Fremont office and beyond. The Berkeley office also maintains a full complement of printing devices, including a high-quality black/white/color copier/printer/scanner, a black/white laser printer, a color ink-jet printer, and a large-format color ink-jet plotter capable of producing E-size output. Drs. Reynolds, Nelson, and Mr. Hertz have offices at the Berkeley location.

Other Resources:

Medical literature: Through its partnership with the Stanford Cancer Institute, CPIC scientists are Cancer Institute members and have access to the core resources available through the University. This includes access to a full array of scientific journals through the Stanford University Libraries. Services available include inter-library loan, reference assistance and journal access, and a pull and copy service (for a small fee). These services are available

through a number of sources including on-line, email and in-person access. CPIC scientists also have access to other core resources through the Stanford Cancer Institute, including access to faculty in the Biostatistics core.

Internet hosting resources: CPIC maintains a comprehensive website that describes our mission, our investigators and our research efforts within communities/populations throughout California, the U.S. and internationally. This website specifically contains a detailed profile of each investigator, all of the research studies and projects as well as a complete list of publications of all investigators. Most of these publications are also linked to PubMed abstracts so that others can view the details of the work conducted by the organization. Additionally, the website has the capabilities to upload education and training materials, certain research instruments (i.e. surveys) and other tools which can be shared with other researchers outside the organization. Confidential data can be transferred using RegWeb, the secured, encrypted cloud-based data sharing system used by the California Cancer Registry.

Regulatory compliance support: CPIC's legal counsel provides investigators with study-specific guidance and support to ensure streamlined IRB review and ongoing compliance with state and federal regulations.

Existing collaborations: Drs. Gomez and Cheng also collaborate with several other investigators in research that is complementary to that proposed in the current application. This study will benefit through critical discussion and feedback as part of regular interactions with these individuals and their study teams. Dr. Gomez also has strong ties to various AAPI community and policy organizations, including the Asian Pacific Islander American Health Forum (APIAHF); Weaving an Islander Network for Cancer Awareness, Research and Training (WINCART); Asian Pacific Partners for Empowerment, Advocacy, and Leadership (APPEAL); Asian Health Services (AHS), and Asian Americans for Community Involvement (AACI). Furthermore, through regular interactions with members and programs within the Stanford Cancer Institute, the study investigators have access to clinical and multidisciplinary expertise that will further enrich this study, as well as provide collaborations for future research stemming off of this proposed study. This table describes some of these useful collaborations (intellectual resources).

Name (Institution)	**Area(s) of expertise**
Loic Le Marchand (Univ of Hawaii)	Lung cancer epidemiology, racial/ethnic differences in cancer risk
Christopher Amos (Dartmouth University)	Lung cancer epidemiology and genetic epidemiology
Anna Wu (USC)	Lung cancer epidemiology, cancer epidemiology in Asian Americans
Beate Ritz (UCLA)	Air pollution assessment and environmental health epidemiology
Vish Nair (Stanford University)	Pulmonary medicine
Judith Prochaska (Stanford University)	Tobacco control research
Myles Cockburn (USC)	GIS, geospatial epidemiology
Lilhua Liu (USC)	Cancer surveillance in Native Hawaiian and Pacific Islander populations
Brenda Hernandez (Univ of Hawaii)	Director, Hawaii Cancer Registry; cancer epidemiology in Native Hawaiians and Pacific Islanders
Janice Tsoh (UCSF)	Tobacco control research
Amar Das (Dartmouth)	Biomedical informatics
Allison Kurian (Stanford University)	Oncology, research with EHR data

STANFORD FACILITIES AND OTHER RESOURCES

Stanford University Medical Center:

Stanford University is a private educational institution located on 8,180 acres, approximately 30 miles south of San Francisco. The School of Medicine has 27 departments (18 clinical, 9 basic and 5 Institutes), with a total of 825 faculty members. In the past year, the School graduated 99 MD students (11 of these were MD/PhDs) and 122 PhD students. Stanford University Hospital has 613 licensed beds and trains over 850 housestaff and fellows. There are three other major affiliated teaching hospitals providing approximately an additional 2000 beds for teaching and research which include the Palo Alto Veterans Administration Health Care System, Santa Clara Valley Medical Center, and Kaiser Permanente Hospital in Santa Clara. The Stanford University Medical Center (SUMC) and the Stanford University Medical School exist as an integrated complex on the University campus. The total laboratory space in the School of Medicine is approximately 360,000 square feet. Total NIH funding for FY 2011 was $334,468,844. The Department of Medicine with its many Divisions is situated in close proximity to the basic science departments within walking distance from each other.

Scientific Environment:

Stanford University provides a unique environment for trans-disciplinary and translational research. The medical center campus houses the Clark Center, a bioengineering hub that links the medical school campus with the engineering campus. This center serves as an anchor for collaborations. Stanford University also has many investigators versed in molecular genetics and genomics that provide a strong environment for development of molecular diagnostics. This expertise and collaboration extends from clinical departments (Thoracic Oncology, Pathology, Surgery, Medicine and the Cancer Center) to basic science departments (Genetics, Statistics, Biochemistry, Developmental Biology).

Clinical:

Dr. Wakelee is the lead thoracic oncologist at Stanford, among a group of 15 oncologists, pulmonologists, radiologists, and pathologists. Dr. Wakelee specializes in treating never smoker lung cancer patients, and sees a large number of Asian American patients, given the regional demographics. She has also successfully recruited patients for a variety of clinical and translational studies for the past decade.

Computer:

Drs. Wakelee, Patel, and Haile have offices in the Stanford Cancer Institute, and high speed computers with networking capabilities. Full 1Gbit networking (internal and external) is managed by Stanford's Medical IT group. Firewalls and system management tools are employed for security and comply with HIPAA. The entire medical center is linked through the Stanford University network (SUNet) which connects over 20,000 computers including mainframes, mini computers, advanced work stations plus several thousand computer terminals. SUNet also provides access to other computers off campus, national super computer centers and computers on the internet throughout the world. Through the internet for many computers on

SUNet, access is granted to a wide variety of biomedical resources such as the National Bibliographic database, Genbank and other molecular biology databases. A number of software programs are licensed by Stanford University which are provided to members of the medical center community. The full range of searching capabilities are readily accessible from a number of terminals throughout the medical center.

Lane Medical Library:

The medical center is supported by the Lane Medical Library. This library is approximately 39,000 square feet, has over 400,000 books and receives approximately 3,000 journals on a regular basis. The library subscribes to over 800 ejournals which are fully accessible at desktop computers. The library provides access to computerized data bases for Melville, Medline, Biological stracts, Toxline, Cancer Lit, AIDsline, PDQ, CINAHL (nursing), Science Citation Index and UnCover. In addition, there are a number of smaller departmental and divisional libraries located throughout the medical center.

Stanford University Core Facilities:

Stanford University has several established core facilities that are available to the research community. All are housed under the Stanford Center for Clinical and Translational Education and Research (SPECTRUM).

Stanford Cancer Institute:

In 2007 the then named Stanford Cancer Center received funding from the NCI, and many Program faculty participate in the currently named Stanford Cancer Institute. There are multiple core facilities that are available to Program faculty and trainees, including an animal tumor models shared resource, bioscience screening facility, biostatistics resource, imaging facility, cell science imaging facility, flow cytometry, genomics shared resource, human immune monitoring core, proteomics core, tissue procurement core, and a clinical protocol and data management core. Institute Members (including Drs. Gomez, Cheng, Reynolds, Nelson, Shariff-Marco, Haile, and Wakelee) have access to the following School of Medicine core facilities relevant to the proposed research:

Biostatistics Facility. Under the direction of the Division of Biostatistics, provides statistical support for School of Medicine research projects. Although Co-Investigator Dr. David Nelson, who is faculty on the Biostatistics Core, will be our primary source of biostatistical support, this core represents an additional resource.

Computational Services and Bioinformatics Facility. Center provides access to a wide array of software resources, including ArcGIS, as well as biomedical informatics expertise.

PALO ALTO MEDICAL FOUNDATION FACILITIES AND OTHER RESOURCES

Palo Alto Medical Foundation Institutional Support

The Palo Alto Medical Foundation (PAMF) (http://www.pamf.org) is multi-specialty, non-profit organization serving patients in Alameda, San Mateo, Santa Clara, and Santa Cruz counties in the San Francisco Bay Area of Northern California, and is the largest group among Sutter Health Affiliates. PAMF patients have a diverse racial and ethnic composition (17% Asian, 11% Hispanic/Latino, 3% African American, 58% non-Hispanic white, and 11% other). All the practice sites have comprehensive electronic health records (EHR) systems and have unified physician payment and billing system.

PAMF provides an excellent venue for patient centered outcomes research due to natural variations in practice across one organization with 1,200+ physicians who are practicing across 70 clinics. PAMF serves patients with a mixture of insurance plans, and yet are linked with a single administrative system.

The Palo Alto Medical Foundation Research Institute (PAMFRI) is a division of PAMF. This relationship provides access to the wealth of data and contextual resources. Researchers at PAFMRI, however, have the same intellectual freedom and independence as those at academic institutions. A critical aspect of several investigators' decision to leave academia and continue their research careers at PAMFRI was PAMF's commitment to supporting cutting-edge health services and policy research, access to data, and intellectual freedom. PAMFRI's transition from a primary focus on bench science to engagement in health services research has been well received by the PAMF clinicians and leadership as they simultaneously seek to lower costs and improve quality of care. PAMFRI researchers participate in various PAMF committees, including the Quality Improvement Steering Committee (QISC), where important issues in service delivery (e.g., practice guideline dissemination) and physician compensation are discussed.

PAMF Clinical Environment

PAMF is staffed by 1,200+ physicians, about half of whom are in primary care. It provides state-of-the-art medical care for about 900,000 patients with over 3.3 million patient visits per year. All physicians and patients are linked through the same EHR, EpicCare. HIPAA compliant, de-identified clinical and demographic data on these patients are readily accessible for research purposes in a secure computing environment. PAMF has been assessed by the Integrated Healthcare Association (IHA) over the years, as have 230-some other physician groups representing approximately 35,000 physicians. Founded in 2001, the IHA P4P program assesses performance using a wide range of HEDIS ambulatory care quality of care indicators, including cardiovascular disease risks.

Only 15% of patient revenues at PAMF are from capitated contracts—the rest are largely covered by various PPOs and Medicare. Thus, the external payment environment is similar to the US medical care system as a whole. With a wide range of payers, there is no single formulary. Because all payers in the area (except for Kaiser) have contracts with PAMF, patients can stay with their physicians when changing jobs. PAMF physician salaries are intended to reflect their own work effort, currently measured by work RVUs. In essence, PAMF incentives for clinicians are similar to what is seen nationally.

PAMFRI Research Environment and Intellectual Rapport

PAMFRI was established in 1950 and has a long history of research, ranging from basic science to health services research. The PAMF Board decided in 2006 to redirect the PAMFRI's focus from lab-based research to investigations that leverage its large EHR database. PAMFRI already had significant ability to support health services and health policy research and is expanding rapidly. It has three senior investigators, two associate investigators, and two assistant investigators, each with major NIH awards. Research activities within PAMFRI generated $5 million in 2014 from federal, industry, and private sources. PAMFRI actively supports multiple externally funded studies on mental health services.

PAMFRI is formally affiliated with neighboring academic institutions (Stanford, UCSF and UC Berkeley), and various other research institutes, yielding active research and training partnerships. It is a NIH Clinical and Translational Science Institute (CTSI) consortium affiliate of both Stanford and UCSF. It is one of the seven cosponsors of a regional research collaborative consisting of PAMFRI, Philip R. Lee Institute of Health Policy Studies (PRL-IHPS) at the University of California, San Francisco (UCSF), Center for Health Policy/Primary Care and Outcome Research (CHP/PCOR) at Stanford, Kaiser Foundation Research Institute, San Francisco Veterans Affairs Medical Center, and Center for Health Research at UC-Berkeley. Jointly with Kaiser Foundation Research Institute and PRL-IHPS at UCSF, PAMFRI is an AHRQ Accelerating Change and Transformation in Organizations and Networks (ACTION II) site. PAMFRI is a member of both the HMO Research Network and AHRQ's Practice Based Research Network.

PAMFRI Computer Support

PAMFRI maintains a local area network comprised of over 90 microcomputers (Macintosh and PC) and 15 servers, including servers running Windows 2012 r2, Windows 2008 r2, and Mac OS, for file sharing, data storage and electronic mail. Appropriate software programs for research needs, including the SAS/JMP and STATA statistical packages, SPSS, Microsoft Office Professional Suite, and Endnote, are installed with regular updates.

The PAMFRI network is protected by a hardware firewall and complex passwords in a domain environment. PAMFRI investigators can also access data in secure PAMF servers maintained by the Information Technology Department, which houses the clinical data warehouse, and data from the Quality and Planning department. All data processing is HIPAA compliant. All research staff have access to full-time technical support for PAMF and PAMFRI networks.

PAMFRI Office

The PAMFRI office building is located on the PAMF main campus (795 El Camino Real, Palo Alto, CA) where the medical and administrative staff offices, conference facilities, as well as the main Palo Alto clinic are located. The PAMFRI facility has approximately 38,000 square feet available for academic offices and staff cubicles, a state-of-the-art focus group facility with a one-way mirrored observation room and video and audio recording capabilities, and 4 fully outfitted patient interview/exam rooms. The office space is equipped with a range of multimedia equipment and smart boards in large and smaller conference rooms, copiers, and fax machines.

PAMF Major Equipment

Access to all necessary equipment for data system (e.g., EpicCare EHR and Solutions® database server) is provided by PAMF at no direct cost.

PAMFRI Other

PAMF has online and on-site resource libraries, which house selected databases and journals on health services and medicine. The investigators at PAMFRI are affiliated with Stanford and/or PRL-IHPS at UCSF. The full resources of the UCSF are therefore available in the support of research, such as access to the entire University of California library system and numerous seminars and training programs. Research materials at the library are available both on campus and through remote access via VPN. Free classes are regularly offered on how to use database software and various search engines to keep investigators updated on the most recent information technology. Librarians are available for consultation to help with searches, systematic reviews, and meta-analyses.

CENTER FOR HEALTH RESEARCH – HAWAII, FACILITIES AND OTHER RESOURCES

KAISER PERMANENTE HAWAII

Located in Honolulu, Hawaii, Kaiser Permanente Hawaii (KPH) was founded in 1958 and is a non-profit, integrated health care delivery system with over 230,000 members, one 285-bed acute care inpatient facility, 20 outpatient clinics on Oahu, Maui, and Hawaii Island, and numerous independent primary care providers on Kauai, Lanai, and Molokai. About 75% of patient membership lives on the island of Oahu, the remainder on Maui and Hawaii Island. Health plan members are highly representative of the state's population, which is one of the world's most ethnically diverse groups (77% minority).

The Hawaii Permanente Medical Group (HPMG) is the state's largest and most experienced multi-specialty medical group practice, made up of over 500 of the most uniquely qualified physicians in the islands and representing over 60 specialties. Named as the highest-rated health insurance plan in Hawaii for commercial, Medicare, and Medicaid lines of business based on quality and member satisfaction, according to the National Committee for Quality Assurance (NCQA), KPH is also the first multi-site health care organization in Hawaii to be recognized by NCQA for Patient-Centered Medical Home (PCMH) Model, with all 20 primary care clinics and providers receiving the highest level of recognition.

CENTER FOR HEALTH RESEARCH, HAWAII

As a joint department of the Kaiser Foundation Health Plan and HPMG, the strategic focus of the Center for Health Research, Hawaii (CHRH) is to advance knowledge to improve health of diverse populations. Formally organized since 1999, CHRH has five doctoral trained scientists, one medically trained scientist, one affiliate investigators (MD), and extensive local and national collaborative arrangements to conduct innovative interdisciplinary scientific research. CHRH focuses on innovative care delivery and translational research uniquely suited to its strong position within a moderately large integrated health care system serving a highly ethnically diverse, defined population. CHRH is a member of the Kaiser Permanente National Research Council, the HMO Research Network, the Cancer Research Network (CRN), and PCORI Patient Outcomes Research to Advance Learning (PORTAL).

Comprehensive electronic databases provide longitudinal information about the demographics, health, and health care utilization of health plan members CHRH provides resources and expertise to support clinical research programs carried out by KPH clinicians. Clinical research is funded by federal, foundation, and proprietary support. CHRH supports clinicians with budgeting, training staff, and other tasks required for fitting research into clinical schedules. All data systems are accessible, with IRB approval of a HIPAA waiver of consent.

FACILITIES

Offices: CHRH is located in a 5,000 sq. ft. space on the second floor of the Kaiser Permanente Dole Consolidated Services Center (CSC) adjacent to the health systems medical records storage facility and the CHRH Clinic. Thus, CHRH has full access to both electronic and paper KPH records with IRB approval of projects. CHRH has access to the hospital library, programmer analysts, medical record technicians, and computer network support.

The CHRH research library maintains access, via Internet and other means, to all national and international databases through Dialog, BRS, and Medline, and can obtain documents from all

major libraries nationwide through DOCLINE and OCLC. This includes the Permanente Knowledge Connection which provides online access to 158 medical journals. High speed Xerography, high speed scanning, photo-quality printers, office equipment, video conferencing equipment, video monitoring and recording equipment, clinical equipment (examination equipment, laboratory equipment), audiovisual equipment, digital recording equipment for telephone interview and focus groups, and personal computers with access to national KP network. Webbased CAPI (Computer-assisted personal interviewing) software and hardware, including touch screen interface.

Laboratory: The CHRH research clinic includes a reception area, 2 exam rooms, 2 office/interview rooms, blood drawing and specimen processing facilities, locked records storage areas, and a large conference room. Local laboratory analyses are performed at the Regional Clinical Laboratory at Moanalua Medical Center.

The 2 office/interview rooms are equipped with independent remote-controlled pan/tilt/focus/zoom cameras and ceiling microphones for live remote audiovisual monitoring with the additional capability of recording sessions on DVD. Both interview rooms can be used and monitored simultaneously. The KPH medical records storage is adjacent to the research clinic.

Clinical: Kaiser Foundation Health Plan, Inc. (KFHP) Hawaii Region is a mixed model Health Maintenance Organization (HMO) serving 230,000 members on the islands of Oahu, Maui, Hawaii, and Kauai. KFHP contracts with Kaiser Foundation Hospital (KFH) for inpatient services and the Hawaii Permanente Medical Group, Inc. (HPMG) for professional services. HPMG is a partnership of 400 physicians, who comprise a wide range of medical specialists and sub-specialists. The Hawaii Region is collaboratively co-managed by KFHP (the insurer), KFH (care facilities), and HPMG (the caregivers). The responsibilities and relationships between these entities are described in this document.

The Hawaii Region provides clinical care services at its own medical clinics on three islands: Hawaii Island (4), Maui (4), and Oahu (10). On Kauai, Molokai, and Lanai KFHP members are cared for in the private offices of a preferred provider network. On Oahu, Kaiser Foundation Hospital is located at the Moanalua Medical Center and is equipped with 285 beds. Additionally, the Hawaii Region contracts for care services with 17 acute care hospitals on all islands for inpatient services. Medically indicated services not available at the Oahu KFH are contracted with community or mainland acute care facilities and practitioners, to ensure available medical care in accordance with the KFHP insurance benefit agreements.

Each Kaiser Permanente member is assigned a unique number upon joining the health plan. This number is retained for life, regardless of leaving and re-joining the plan.

COMPUTER AND DATA MANAGEMENT RESOURCES

Computing resources: CHRH has a network computer system with access to KP national consisting of PCs, servers, and workstations running on different platforms. Every investigator in CHRH is furnished adequate office space with telephone service and a current computer with necessary word processing, statistical, database, e-mail, and spreadsheet programs. Investigators also receive a Hewlett Packard laptop equipped with the technical capabilities necessary to conduct research remotely and securely. Additionally, other major equipment at CHRH include wireless network connections, high speed Xerography, high speed scanning, photo-quality printers, video conferencing equipment, video monitoring and recording equipment, clinical equipment (examination equipment, laboratory equipment), audiovisual

equipment, digital recording equipment for telephone interview and focus groups. All computer workstations and servers are protected behind firewalls. CHRH manages and analyzes large-scale data collections, and has the capacity to support and provide multi-faceted, internet-based services through the use of both public and secure private external network connections.

Data Systems: KPH utilizes a variety of operational systems to support its business and clinical functions. These data systems pertain to a number of related subject areas: enrollment/eligibility; inpatient and outpatient encounters; case management; financial reporting; membership cards; invoices and reconciliation. The major systems are described briefly below.

KP HealthConnect: In December 2004 Kaiser Permanente began implementing a new integrated electronic medical record (EMR) 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 prolonged hospital admission. The standard Epic products have been extensively customized for the needs of KP where it is now referred to as KP HealthConnect (KPHC). The new 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. Other functions, including labs, radiology, oncology and other diagnostic testing, tracking of outside claims and referrals, and membership enrollment, continue to depend on legacy applications. In addition to replacing existing legacy functions, KPHC has enhanced significantly the scope and detail of available data.

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. The KP HealthConnect workstation is located in each exam room accessible to over 3,000 staff. Though KPHC was implemented in 2004/2005, KPH has captured inpatient diagnoses and procedures electronically since 1985 and pharmacy, laboratory tests and claims since 1995. KPHC documents all outpatient and inpatient encounters, billing, appointments, patient registration, surgery operating room management, provides the MyChart Web portal for members, and serves as the clinical data source for the new Enterprise Data Warehouse. The data captured by KPHC is accessible for analysis and includes both coded fields and free text notes. CHRH uses natural language processing algorithms to assess the content of free text and to incorporate it into data analysis.

Every KPHI 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 medical record. Access to the data by researchers 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 a 1-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 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 HC applications, and provide information on patient outcomes, clinical effectiveness, and quality. This provides a data base platform for health services research and invaluable data for epidemiologic studies that reflect actual patient care.

Decision Support System (DSS) draws from all management information systems and provides integrated reports to appropriate managers on daily, monthly, or as needed basis. It integrates department general ledger data with KP HealthConnect to provide inpatient and outpatient activity data. In addition to supporting budget and actual variance monitoring, DSS serves as the major reporting database for utilization/financial reporting, cost analysis, Medicare ACR, trending, and other regional informational needs.

The Inter-Regional Common Membership System (CMS) is an integrated membership enrollment and eligibility system which stores benefits and rates for group account contracts. It tracks billing information and reconciles group and individual dues payments; assigns medical record numbers and issues member identification cards. It interfaces with other operational systems to provide member data and provides a membership reporting database for forecasting, decision support, and other regional information needs.

The Kaiser Permanente Outside Purchases System (KPOPS) tracks all claims and referrals and pays outside providers for health care delivered outside KPH's service system. Encounter data for referrals are linked to DSS.

Laboratory Information System (LIS) stores all laboratory and pathology orders and results.

Radiology Information System (RIS) identifies and tracks all KPH radiology and imaging services.

Pharmacy System (RXIS) tracks pharmacy utilization (including costs) associated with dispensing drugs and prescriptions written at KPH facilities. Pharmacies are located in each of the 17 clinics and in the Moanalua Medical Center. The automated system works in tandem with KP HealthConnect to record all medication orders (inpatient and outpatient). It also tracks all prescription dispensing and refills. The system has been in operation since 1987 and contains data on over 20 million dispensing orders. More than two million dispensing orders occur each

year. Approximately 90% of the KP Hawaii members have drug benefits and about 95% of those members fill their prescriptions at a KP facility. Standard nomenclature is used to identify the individual medications (NDC and GPI codes), along with route (oral vs. inhaled etc), dose, quantity, days supply, provider and member identification. The data are extracted and loaded each night into the central KP Hawaii data warehouse.

Panel Support Tool (PST) - KPH originated the panel support tool, a feedback mechanism for providers based on KP HealthConnect. This tool summarizes 32 quality measures for major chronic diseases and identifies persons with those diagnoses. The PST shows detailed data on member status, chronic disease status, service deficiencies and provides other helpful flags to effectively manage a large panel of patients. The PST is used as the primary source for determining patient eligibility for WellRx.

Tumor registry (Precis) - KPH has over 15,000 cancer cases in a database initiated in 1960. In 1988, the Regional Tumor Registry began, extending registry coverage beyond the island of Oahu to all islands in Hawaii. In Hawaii, all cases of malignancy found by hospitals and nursing homes are reportable by law to the Hawaii State Tumor Registry. The Tumor Registry identifies patients who are eligible for cancer clinical trials and also identifies patients who have missed follow-up appointments. Hawaii's Tumor Registry is part of the National Surveillance Epidemiology and End Results (SEER) Program sponsored by the National Cancer Institute. In addition to being part of the Hawaii SEER program, KPH's cancer program has been approved by the American College of Surgeons since 1977. Data is sent to the American College of Surgeons annually for inclusion in the National Cancer Database. The local database is computerized using Precis software and is available to Kaiser Permanente physicians for studies and research papers. KPH has also been affiliated with the University of Hawaii's Cancer Research Center of Hawaii (CRCH) for more than 20 years.

VIRTUAL DATA WAREHOUSE (VDW)

The VDW is a collaborative effort of the HMO Research Network (HMORN), a consortium of research groups in 18 health plans across the country. Each HMORN health plan maintains a VDW as a platform to facilitate multi-site research. The VDW is a series of dataset standards and automated processes that allow programs written at one HMORN Site to be run against other VDW sites quickly and with a minimum of site-specific customization.

The VDW is "virtual" in the sense that data remain at each site; the VDW is 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" procedure produces SAS datasets with a common structure at each HMORN site, thereby allowing an analyst at one site to write one SAS program to extract and/or analyze data at all participating sites with minimal edits.

VDW Datasets/tables cover the following concept areas:

- **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. As many records as are necessary are added to represent changes over time in, say, insurance type(s), or completeness of data capture.

- **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.
- **Tumors**: 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.
- **Demographics:** Birthdate, gender, race, and ethnicity (as available) for everyone enrolled or ever treated at the health plan.
- **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 HMORN and CHRH investigator interest, the VDW Vital Signs table also includes anthropometry (height, weight) and tobacco use.
- **Death:** Contains death date and cause of death information.
- **Census Demographics:** Contains 2000 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.
- **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.

In addition to maintaining SAS datasets as a version of the VDW, CHRH also maintains an instance of the VDW within CHRH Research Database in Oracle database format. CHRH VDW contains many local variables requested by CHRH programmers and investigators, in addition to the standard HMORN variables.

Overall, the VDW presents a rich data resource for research applications. Data are harmonized across various data sources. In the case of KPNC, data that are contained 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 HMORN settings, the ability to conduct collaborative research across these

systems is enhanced substantially, and thus, provide an outstanding resource for collaborative studies, both within KPNC alone and across the HMORN.

PHS 398 Cover Page Supplement

OMB Number: 0925-0001

1. Project Director / Principal Investigator (PD/PI)

Prefix:	Dr.
First Name:	Scarlett
Middle Name:	Lin
Last Name:	Gomez
Suffix:	Ph.D.

2. Human Subjects

Clinical Trial?	☒ No	☐ Yes
Agency-Defined Phase III Clinical Trial?*	☐ No	☐ Yes

5. Human Embryonic Stem Cells

* Does the proposed project involve human embryonic stem cells?	☒ No	☐ Yes

PHS 398 Research Plan

OMB Number: 0925-0001

Please attach applicable sections of the research plan, below.

1. Introduction to Application (for RESUBMISSION or REVISION only)	
2. Specific Aims	1255-Specific_Aims.pdf
3. Research Strategy*	1256-Research_Strategy.pdf
4. Progress Report Publication List	
Human Subjects Sections	
5. Protection of Human Subjects	1257-Protection_of_Human_Subjectcs.pdf
6. Inclusion of Women and Minorities	1258-Inclusion_of_Women_and_Minorities.pdf
7. Inclusion of Children	1259-Inclusion_of_Children.pdf
Other Research Plan Sections	
8. Vertebrate Animals	
9. Select Agent Research	
10. Multiple PD/PI Leadership Plan	1260-Multiple_PD_PI_Leadership_Plan.pdf
11. Consortium/Contractual Agreements	1261-Consortium_Contractual_Arrangements.pdf
12. Letters of Support	1262-Letters_of_Support.pdf
13. Resource Sharing Plan(s)	
Appendix (if applicable)	
14. Appendix	

SPECIFIC AIMS

Lung cancer is the leading cause of cancer deaths among Asian Americans, Native Hawaiians and Pacific Islanders (AANHPI), a diverse population in the U.S. that includes individuals from more than 30 different countries, speaking more than 100 languages. For females of specific AANHPI ethnic groups, lung cancer is the 3rd or 4th most common cancer, but the most common cause of cancer death. The burden of lung cancer among AANHPI females is striking considering their low prevalence of smoking. Risk factors beyond smoking remain largely unknown. Among AANHPI females, approximately 70% of lung cancers occur in never smokers. Furthermore, the incidence rates of lung cancer, especially adenocarcinoma, are either stable or increasing among Filipina, Korean, and Chinese American females, in stark contrast to the overall declining incidence rates of lung cancer in other U.S. racial/ethnic groups. Population-based registry data from the Surveillance, Epidemiology, and End Results (SEER) program have been used to document lung cancer incidence trends among racial/ethnic groups. However, given the lack of information on smoking status for populations at risk, population-level incidence rates stratified by smoking, race/ethnicity, and gender are not available, and constitute a critical gap in knowledge. At present, there is no single sufficiently large data source to document lung cancer incidence rates by smoking status among specific AANHPI ethnic groups, which is central to understanding and reducing the burden of disease in this heterogeneous population.

Based primarily on data from Asia, the epidemiology of lung cancer appears to be distinct among AANHPI females, with proportionally higher incidence of adenocarcinoma, the histology least associated with smoking. Furthermore, lung cancers among AANHPI females are more likely to have EGFR-TK mutations and EML4- ALK translocations. Studies in Asia have also identified several risk factors for lung cancer among East Asian female never smokers, including second-hand tobacco smoke (passive smoking), exposure to indoor cooking oil fumes, and outdoor air pollution. Body size and reproductive factors such as use of hormone therapies may also have etiologic significance, but thus far, results from studies have been mixed. However, together these risk factors only explain a small percentage of lung cancer risk among females in Asia and the prevalence of most of these exposures are considerably lower among AANHPIs in the U.S. Moreover, etiologic studies of lung cancer risk among AANHPIs have been sparse and limited to older case-control studies focused mainly on active smoking history. **Without a clear understanding of the etiologic factors underlying lung cancer risk among AANHPI females, we are at a loss for designing the appropriate intervention strategies to reduce their high burden of disease.** A focused study on risk factors in such populations may provide valuable insights into the etiology of lung cancer among never smokers in general, which is believed to be a separate disease from lung cancer due to smoking.

The objective of our study is to leverage prospective data from two large electronic health record (EHR) databases, comprising 3.2 million individuals, 1.8 million females, and over 240,000 AANHPI females, with up to 15 years follow-up, to estimate their lung cancer incidence and characterize the epidemiology of lung cancer by specific single and mixed race/ethnicity and smoking status. Our specific aims are as follows:

Aim 1. To characterize lung cancer incidence rates among AANHPI females by specific single and mixed race/ethnicity groups and smoking status. We will calculate overall and histological cell-type specific incidence rates of lung cancer by smoking status. We will also compare the distribution of sociodemographic, tumor (e.g. stage, histology), and molecular characteristics (e.g. EGFR, ALK) of lung cancer cases by race/ethnicity and smoking status. Males and other races/ethnicities will be examined for comparison.

Aim 2. To identify multi-level etiologic risk factors for lung cancer among female AANHPI never smokers. We will conduct a longitudinal analysis of lung cancer risk, including absolute risk modeling, examining six exposure domains: second-hand smoke, previous lung diseases, infections, reproductive history and hormone exposure, body size, and neighborhood environmental factors, including measures of particulate matter (PM), traffic density, neighborhood socioeconomic status (nSES), and ethnic enclave.

This study will use EHR data from the Northern California Sutter Health system and from Kaiser Permanente Hawaii – each specifically selected for their robust AANHPI representation and high quality data. Coupled with a focus on distinct subpopulations defined by race/ethnicity (including well-defined mixed racial/ethnic groups), smoking status, environmental characteristics, and tumor-based molecular markers, this study will facilitate NCI's goal of precision medicine - "prevention and treatment strategies that take individual variability into account" (Collins and Varmus, NEJM 2015) – which is particularly suitable for large-scale EHR-based studies. This highly efficient study will have the unprecedented capability to provide the much-needed information on lung cancer risk among AANHPI never smokers, serving as a critical evidence base to inform screening, research, and public health priorities in this growing population.

RESEARCH STRATEGY

A. SIGNIFICANCE

A1. AANHPIs are heterogeneous and the fastest growing racial/ethnic population in the U.S., increasing by 46% between 2000-2010 [1], and surpassing the number of Hispanic immigrants for the first time [2]. Our study sites, Hawaii and Northern California, comprise among the nation's largest populations of AANHPIs [3, 4]. Encompassing people from 30 countries, speaking more than 100 languages, AANHPIs are one of the most heterogeneous racial/ethnic groups in the U.S., with great diversity in socioeconomic levels, cultural beliefs and behaviors, degree of English proficiency, immigration experience, generational status, and acculturation [5-8]. For example, median household income varies from $50,000/year among Koreans to $88,000/year among Asian Indians; 70% of Asian Indians have a college degree in contrast to 26% of Vietnamese [5]. AANHPIs are also unique demographically in terms of persons identifying as mixed or multiple races, populations that have increased 60% among AAs and 44% among NHPIs from 2000 to 2010 [3, 4]. Despite this great diversity, research historically has considered AANHPIs as one aggregate group, consequently masking health disparities that may exist among distinct ethnicities. This approach has unfortunately led to the long-held misperception that Asian Americans are a "model minority", enjoying relative affluence and good health [5]. With more recent research now focusing on the disaggregation of this population, there is emerging evidence of the substantial health and exposure inequities that debunk this perception of homogeneity [6, 7, 9]. Consistent with recent U.S. Department of Health and Human Service guidelines to report data for distinct AANHPI groups, a notable strength of this proposal is the ability to evaluate multiple specific AANHPI groups and well-defined AANHPI mixed races/ethnicities (i.e., Asian and NHPI, Asian and White, NHPI and White).

A2. Current data on lung cancer incidence rates among AANHPI females by smoking status are lacking. The heterogeneity across AANHPI ethnic groups is reflected in lung cancer rates and trends across these specific groups. Using data from 13 SEER registries, we found incidence rates from 1990-2008 to vary across eight AA groups, from 12.4 per 100,000 for Asian Indian and Pakistani females to 31.8 for Vietnamese females. Moreover, while lung cancer incidence rates have been decreasing for non-Hispanic White females, a 2.1% annual increase from 1990-2008 was seen for Filipino and Korean females [10]. In a companion publication, we found that Native Hawaiian and Samoan females had similar incidence rates as non-Hispanic White females, while rates for Guamanian/Chamorro females were lower [11]. These studies, while demonstrating variability in lung cancer incidence rates across AANHPI groups, were not able to report rates by smoking status given the lack of smoking data in the numerator and denominator.

Although smoking is the most important cause of lung cancer, unlike AANHPI males, AANHPI females (with the exception of Native Hawaiians [8, 12]) have a low prevalence of smoking. In 2009-2011 data from the California Health Interview Survey, the prevalence of current smoking ranged from 1% for Chinese females to ~7% for Japanese, Korean, and Filipino females. In 2009-2010 data from the National Adult Tobacco Survey, 2.1% of AANHPI females ever smoked, with the highest prevalence of 8.3% among Japanese [12].

A3. Lung cancer is more common among female than male never smokers, with a potentially greater disparity among AANHPIs. In a highly-cited publication, we estimated incidence rates of non-small cell lung cancer (NSCLC) by sex and smoking status using data from several large prospective cohorts [13], largely consisting of non-Hispanic Whites. Incidence rates ranged in these cohorts from 15.2-20.8 per 100,000 person-years among never-smoking

females and 4.8-13.7 among never-smoking males; however, a limitation of these data was the inability to estimate reliable rates by race/ethnicity. In a population-based series of female NSCLC cases in the San Francisco Bay Area, 70% of AANHPI cases were never smokers, in contrast to 35% of Latina and 10% of non-Hispanic Whites cases [14]. In Asia, at least half of lung cancers among females occur among never smokers [15]. In an early study of Chinese, Japanese, Hawaiians, and Filipinos in Hawaii; never-smoking Chinese females had increased lung cancer risk compared to other never-smoking females [16]. The importance of considering smoking status among AANHPI females has also been suggested by histological cell-type specific incidence rates. Recently, we reported that the incidence trends of adenocarcinoma, the more prevalent histology among never smokers [17], increased from 1990-2010 for Filipino and Korean females, while rates among non-Hispanic White females decreased. In the same period, rates of histologies more strongly associated with smoking - squamous cell carcinoma, small cell, and large cell or other small cell - generally declined or remained stable for all AANHPI groups [18]. By providing contemporary incidence data by smoking status for AANHPI ethnic groups, this study fills a critical gap in knowledge that has limited our ability to quantify the burden of this disease among the large population of AANHPI females.

A4. Lung cancer among never smokers is increasingly believed to be a different disease from that in smokers. Adenocarcinoma is less strongly associated with tobacco smoke than small cell and squamous cell subtypes. In a meta-analysis of case-control studies from Europe and Canada, the odds ratio (OR) for adenocarcinoma comparing heavy smokers to never smokers was 21.9 (95% CI = 16.6 to 29.0) among males and 16.8 (9.2 to 30.6) among females, relative to 111.3 (69.8 to 177.5) for small cell lung cancer in males and 108.6 (50.7 to 232.8) for females, and 103.5 (74.8 to 143.2) for squamous cell lung cancer in males and 62.7 (31.5 to 124.6) for females [17]. Among never smokers, lung cancer is also more likely to target distal rather than proximal airways [19]. Furthermore, lung cancers among never smokers show different mutational patterns compared to that among smokers; with a lower mutational load [20], a greater frequency of EGFR mutations [19] and EML4-ALK translocations [21, 22], a lower frequency of KRAS and TP53 mutations [19], and differential DNA methylation patterns [19]. Moreover, the considerably higher frequency of adenocarcinoma among AANHPI females - more than 45% in most female AANHPI ethnic groups except Native Hawaiians, compared to 35% among non-Hispanic White females and AANHPI males, and 31% in non-Hispanic White males [18] – points to a risk factor profile independent of smoking and perhaps distinct from other racial/ethnic groups and males. Thus, research of etiologic factors should evaluate never smokers separately from smokers, in addition to considering differences by sex and race/ethnicity.

A5. Epidemiologic data on risk factors for lung cancer in AANHPI female never smokers are scant. Exposures that have been associated with lung cancer in never smokers include second-hand smoke [23], family history of lung cancer [24], radon, coal for household cooking and heating, and environmental air pollution [25]. Additional putative risk factors include history of lung diseases [26], reproductive factors and hormone exposure [27-39], body size [40, 41], and infections [19]. However, results from studies on these putative risk factors have been mixed, likely because they have not always considered these associations with attention to sex, race/ethnicity, and histology, or among large populations of never smokers, which likely differ in their effects. Subramanian and Govindan in their review of lung cancer in never smokers summarize that the relative contribution of these established and putative risk factors to lung cancer risk among never smokers has not been well characterized in different racial/ethnic populations [42]. Using the rich clinical data available from EHRs as well as our well-curated data on environmental exposures, our proposal will provide the unprecedented opportunity to provide insights into the relative contribution of six exposure domains to lung cancer incidence among AANHPI female never smokers: 1) second-hand smoke; 2) previous lung diseases; 3) infections; 4) reproductive

history and hormone exposure; 5) body size; and 6) environmental factors including PM, traffic density, nSES, and ethnic enclave. We briefly review pertinent literature for each of these six exposure domains below.

A5a. Second-hand smoke. Second-hand smoke is an established risk factor for lung cancer in never smokers and may be more relevant to females than males, but the degree to which it contributes to the burden of lung cancer in never-smoking females is unclear [43-47]. Estimates of the proportion of lung cancers among never-smoking males and females attributable to second-hand smoke range from 15%-35% [38], with one study reporting 37% among Taiwanese female never smokers [48]. However, estimates for U.S. female never smokers are not available. Interestingly, a recent genome-wide study of mutational profiles among Asians reported that the mutational profiles of female never smokers did not resemble that of smokers, suggesting that lung cancer among Asian female never smokers was not the result of second-hand smoke or other carcinogens that cause oxidative DNA damage as seen in a smoking-associated mutational profile [20].

A5b. Previous lung diseases. Retrospective case-control studies have implicated previous lung diseases (asthma, chronic bronchitis, emphysema, chronic obstructive pulmonary disease, and idiopathic pulmonary disease) as risk factors for lung cancer among never smokers, but results differ by study [26, 44, 49-52]. The most established of these is likely the effect of idiopathic pulmonary disease, but the degree of confounding of this association by smoking is unknown 53. In addition, asthma was shown in a meta-analysis to be positively associated with lung cancer risk independent of smoking status [54].

A5c. Infectious disease. Certain infectious diseases, specifically pneumonia, pulmonary tuberculosis (TB), Chlamydia, and human papillomavirus (HPV), may contribute to lung cancer risk. Results for pneumonia have been mixed [26, 52, 55]. A meta-analysis of 13 studies reported a positive association of TB with lung cancer risk among never smokers (OR=1.78, 95% CI: 1.42-2.23); this association persisted among Asian and non-Asian study populations [56]. The infection rate of TB is particularly high among AANHPIs, especially foreign-born AAs. In 2008, the incidence rate of TB was 23 and 14 times higher among AAs (25.6 per 100,000) and NHPIs (15.9), respectively, than non-Hispanic Whites (1.1) [57]; accordingly, TB infections may be a relevant risk factor among never-smoking AANHPIs. A meta-analysis of 12 case-control studies confirmed an association of seropositive Chlamydia IgA titers with lung cancer risk (OR=1.48; 95% CI: 1.32-1.67) [58]. In addition, in a nested case-control study from the Prostate, Lung, Colorectal and Ovarian cancer screening trial, greater risk of lung cancer among individuals positive for Chlamydia did not differ by smoking status [59]. The role of HPV in lung cancer is controversial. In studies testing for HPV in lung tumor tissue, HPV infection in lung cancer is more prevalent among female never smokers, but epidemiologic data are not available [60]. However, as a greater percentage of Asian lung cancers are HPV positive (35.7% of lung cancers in Asia compared to 17% in Europe and 15% in the U.S.) [60], it may be an important risk factor among never-smoking AANHPIs.

A5d. Reproductive history and hormone exposure. The association of reproductive history with lung cancer incidence is unclear, but laboratory studies indicate that estrogen receptor-beta is more frequently expressed in lung cancers of never smokers [38]. Three cohort studies in Asia considered these factors among never smokers and found associations of age at menarche, parity, length of reproductive period, oral contraceptive use, and age at menopause with lung cancer incidence, but results for these factors are qualitatively different across the studies [29-31]. In addition, all three of these studies reported no association of hormone replacement therapy with lung cancer risk while two studies from the U.S. (adjusted for smoking status) report a

reduced risk [29-32, 39]. An association of reproductive factors and hormone exposure with lung cancer risk may not be straight-forward and studies among never smokers in the U.S. are lacking, thus an investigation of these factors among AANHPI female never smokers in a large U.S. cohort is timely.

A5e. Body size. Previously reported inverse associations of body mass index (BMI) and lung cancer risk have been attributed to confounding by smoking and pre-existing lung disease. BMI results in never smokers are mixed, with most studies reporting no association with lung cancer in never smokers [61-63]; however, one case-control study and one cohort study reported significant associations of increasing BMI with lung cancer risk [40, 41]. BMI will be a particularly interesting risk factor to explore among AANHPIs given that cardio-metabolic diseases manifest among AA at lower BMI thresholds 64, 65 and NHPIs present with higher levels of BMI [6, 8].

A5f. Environmental factors. In 2013, IARC classified outdoor (ambient) air pollution and PM as carcinogenic to humans (Group 1) and a cause of lung cancer [66]. On the basis of 14 studies, the meta-relative risk for lung cancer per 10 ug/m3 increase of $PM_{2.5}$ and PM_{10} was 1.09 (95% CI: 1.04-1.14) and 1.08 (95% CI: 1.00-1.17), respectively [67]. Characteristics of the groups most susceptible to the effects of air pollution are poorly understood [67, 68]. A meta-analysis has shown differences in associations between $PM_{2.5}$ and lung cancer by smoking status with more pronounced associations for former smokers (HR=1.44, 95% CI: 1.04-2.01), intermediate for never smokers (HR=1.18, 95% CI: 1.00-1.39), and weaker for current smokers (HR=1.06, 95% CI: 0.97-1.15) [67]. In a Japanese cohort study, lung cancer risk increased significantly in relation to $PM_{2.5}$ among female never smokers (RR=1.16; 95% CI: 1.02-1.33) [69]. In preliminary analysis, using data from the Multiethnic Cohort, an increased risk of lung cancer was observed with a 1000 ug/m^3 increase in PM_{10} among female Japanese American never smokers in California (HR=2.42; 95% CI: 1.16-5.05). Air pollution may be particularly important for AANHPIs as there is evidence documenting high levels of ambient air pollution exposures, including high traffic volume [70] and $PM_{2.5}$ [71, 72], among AANHPI populations in certain areas [70-80]. AANHPIs are more likely than other racial/ethnic groups to live in counties that exceeded the 24-hr health standards for $PM_{2.5}$ data [72]. We recognize the complexity of assessing air pollution in the Hawaiian Islands due to volcanic smog and fog ("vog"), and thus will limit our PM analyses to the northern California sample.

Emerging research shows that individuals' health can be determined as much by their neighborhood context as by their individual-level behaviors or characteristics [81-85]. Neighborhood social factors such as SES and ethnic enclaves may moderate the impact of air pollution measures and individual-level factors on lung cancer risk. Ethnic enclaves, defined as geographic areas with distinct concentrations of ethnic and/or immigrant groups and cultural mores, are often in areas that are impacted by urban and industrial developments that influence the ambient air in these communities. At the same time, these communities may have greater availability of coethnic social support, improved built environments (due to high density, mixed-use), and access to healthy and affordable ethnic food that may moderate exposures to environmental pollutants [76-78]. However, research on the independent and joint effects of neighborhood- and individual-level factors and lung cancer risk is absent. There is limited but provocative support for the current project's aim to better understand how PM levels, traffic density, ethnic enclaves, and nSES impact lung cancer incidence among AANHPI populations.

A6. Ongoing studies of lung cancer and AANHPIs. NIH Reporter shows one currently funded study, a P50 project (PI: J. Ayanian) from the Centers for Population Health and Health Disparities, that is focused on lung cancer outcomes among racial/ethnic groups in the Cancer Care Outcomes Research and Surveillance (CanCORS) Consortium; however, this project is

focused on lung cancer outcomes (not etiology) and has insufficient numbers of AANHPIs to disaggregate into distinct subgroups. Moreover, in a recent review of the NCI grant portfolio of cancer studies among AANHPIs, Nguyen et al. found no recently funded grants focused on cancer etiology [86]. Clearly, there is large gap in our understanding of lung cancer risk among the fastest-growing racial/ethnic group in the U.S.

A7. Impact. As the number one cause of cancer deaths, lung cancer presents a significant disease burden, which is a particular concern for AANHPI females given their exceedingly low smoking prevalence, lack of data on incidence rates by smoking status for distinct ethnicities, and limited available research regarding risk factors. Our proposed EHR-based cohort study will be the first study to characterize the burden of lung cancer among distinct AANHPI racial/ethnic groups according to smoking status. With roughly 71,000 Japanese Americans and Native Hawaiians in the on-going Multiethnic Cohort Study [87], our proposed resource of more than 418,000 AANHPI females and males will provide the unprecedented opportunity to examine other AANHPI ethnicities not well represented in the Multiethnic Cohort. Leveraging the rich clinical EHR information and well-curated contextual data on air pollution measures, traffic density, neighborhood SES, and ethnic enclaves, this study will additionally evaluate novel lung cancer risk factors and characterize molecular features of lung cancer among AANHPI female never smokers. **Our proposal will thus provide the much-needed information to inform screening, research, and public health priorities in the growing populations of AANHPI.**

B. INNOVATION

Overcoming the challenges in quantifying the burden of lung cancer in relation to smoking behavior and identifying risk factors for disease among specific AANHPI ethnicities requires a rigorously designed epidemiologic study with an extremely large study population. This proposal incorporates several innovative elements with respect to study design and statistical methods to carry out a large-scale study of lung cancer incidence among AANHPIs with an emphasis on never-smoking females. Our proposal is innovative for the following reasons: **(1)** it efficiently leverages existing resources of EHR and complex multi-level datasets to integrate demographic, health behavior, and clinical data collected from two patient-centered health plans from California and Hawaii, states with the largest AANHPI populations, with neighborhood environmental and surveillance data; thus, creating a whole that is greater than the sum of its parts for 3.2 million subjects (1.8 million females, 1.4 million males); **(2)** it capitalizes on routine clinical data that were collected prospectively for a longitudinal analysis that minimizes differential misclassification, survival bias, and selection bias inherent in retrospective designs; **(3)** it recognizes the heterogeneity in AANHPIs and includes highly granular race/ethnicity EHR data to disaggregate this group into specific AANHPI ethnicities [88]; **(4)** it addresses the demographic shift in the growing numbers of people of mixed race/ethnicity by comparing incidence patterns and risk associations across well-defined mixed AANHPI populations; **(5)** it moves beyond the traditional single-level risk model to a multi-level risk model that examines data from several levels of inference, allowing for testing the effects at a macro-level (e.g., neighborhood differences) while controlling for potential confounding at another level (e.g., individual-level factors) and examining the interactions between environment and individual level factors [89, 90], thus, comprehensively integrating information across different systems, and moving us closer to the model of precision medicine [91]; **(6)** it incorporates absolute risk models to quantify the probability of disease development for a given exposure; and **(7)** it includes a molecular epidemiology component to characterize the epidemiologic architecture of key molecular markers such as EGFR mutations that inform precision medicine for clinical decision making for lung cancer. To our knowledge, this will be the first large-scale study of lung cancer

incidence among AANHPI ethnicities that will identify risk factors for these understudied and rapidly growing U.S. populations.

C. APPROACH

C1. Justification and Feasibility

C1a. Study team. This work is led by a multi-disciplinary team with experience in cancer surveillance, particularly among AANHPI ethnic groups, lung cancer epidemiology and oncology, research using EHR data, and environmental epidemiology. Multiple PIs, **Drs. Scarlett Lin Gomez** and **Iona Cheng,** bring complementary expertise in social (Gomez) and molecular (Cheng) epidemiology, and both have expertise in research in the areas of cancer surveillance, lung cancer, neighborhood contextual factors, and on cancer epidemiology among AANHPIs [2, 10, 11, 14, 18, 85, 92-100]. Co-Investigators **Drs. Peggy Reynolds** and **Salma Shariff-Marco** contribute expertise in environmental epidemiology, specifically exposure assessments of air pollution measures (Reynolds) and social and built environment measures (Shariff-Marco). **Dr. Mindy DeRouen** is an early-stage investigator with training in cancer biology and epidemiology; serving as Co-Investigator on this project will provide her with training on a project that brings these components together and positions her to develop future research ideas. Biostatistician **Dr. David Nelson** has extensive experience working with longitudinal and environmental data including application of multi-level and geospatial modeling techniques. In addition to facilitating access to the EHR data, Co-Investigators, **Dr. Beth Waitzfelder** from the Kaiser Permanente Hawaii Center for Health Research, and **Drs. Caroline Thompson** and **Hal Luft** from Palo Alto Medical Foundation Research Institute (PAMFRI), bring critical expertise in the use and interpretation of their EHR data. Dr. Thompson additionally brings expertise in data harmonization and methods for quantitative bias modeling, and Dr. Luft brings expertise in the application of EHR data for addressing research questions. Dr. Waitzfelder additionally brings insights into the social determinants of health with particular knowledge about AANHPI populations. **Dr. Heather Wakelee** is the lead thoracic medical oncologist and co-leader of the thoracic oncology clinical research group at Stanford University; she has authored several seminal papers on lung cancer among never smokers. **Dr. Manali Patel** is an Instructor in Oncology who brings perspectives in health services research. **Dr. Robert Haile** is the Associate Director of Population Sciences at the Stanford Cancer Institute and is an internationally-recognized cancer and genetic epidemiologist, and, of particular relevance to this study, an expert in trans-disciplinary, team-science research. This team has a successful history of collaborations with pilot studies and publications, coalescing into this proposed study. We anticipate that this study will provide the foundation for future collaborative work on understanding the contributing factors to lung cancer incidence and mortality focused on never smokers, females, and disaggregated racial/ethnic populations. Described further here are our relevant previous and current studies that demonstrate our expertise and support the feasibility of the proposed work in our hands.

C1b. Lung cancer epidemiology. Our team has collaboratively published 13 epidemiologic manuscripts on lung cancer incidence and survival. These include: documentation of incidence patterns among AA ethnic groups by nativity [101]; data from four cohorts on incidence by sex and smoking status showing higher incidence among female than male never smokers [13]; an invited commentary on sex differences in lung cancer susceptibility [102]; publication of first-ever data on 19-year incidence trends among AANHPI ethnic groups [10, 11]; factors associated with survival among AANHPIs 103, and among AANHPI and Latina female never smokers [14]; publication of 20-year incidence trends by histology among AANHPI ethnic groups [18]; clarifications about incidence patterns among Hispanics by nativity, and by histology [94, 104]; social factors associated

with survival among Hispanics [105]; role of active and passive smoking in lung cancer incidence in the Women's Health Initiative [106]; and trends of histologic lung cancer incidence among Chinese Americans by nativity [107].

In two manuscripts currently under development, we are evaluating associations of nSES with histologic-specific incidence rates of lung cancer across racial/ethnic groups. Among AANHPI females (aggregated due to data limitations), we find lower nSES associated with higher incidence rates of overall lung cancer but differences in associations across histologic cell types. For small cell and squamous cell lung cancer, lower nSES was associated with higher incidence rates, while adenocarcinoma incidence was not associated with nSES. The variability in observed nSES associations provide strong support for taking the next step in examining both nSES and individual level factors jointly, incorporating additional contextual factors, and disaggregating AANHPIs into specific populations.

C1c. Etiologic risk factors and lung cancer risk in females who never smoked. Co-Investigator **Dr. Reynolds** was the PI for the Bay Area's participation in one of the largest and most comprehensive studies of second-hand smoke (SHS) and lung cancer incidence in lifetime non-smoking females. This much cited multi-center study found a 30% increased risk for lung cancer among females who had never smoked and who reported SHS exposure from a spouse, other household member, co-worker, or in social settings [108-110]. These findings, and evidence for significant dose-response patterns, were reinforced in a pooled analysis with a European multi-center study [111]. The U.S. study also found significantly elevated risks for lung cancer in females with previous non-malignant lung diseases, particularly asthma and chronic bronchitis, independent of SHS exposure [26]. These studies did not include adequate representation of AANHPI females to enable ethnic-specific analysis, providing strong justification for evaluating these risk factors – SHS and lung diseases – among AANHPI females.

C1d. Air pollution exposure assessment and research. **Drs. Reynolds and Nelson** have conducted numerous GIS-based epidemiologic health studies of ambient air pollution, focused on a variety of health endpoints, including cancer, cardiorespiratory disease, and mortality [112-122]. As part of these investigations, this team has devoted considerable effort towards developing and evaluating methodological strategies for estimating population-level and individual-level exposures to ambient air pollution in California [70, 112, 123]. In addition, **Drs. Cheng, Shariff-Marco, and Gomez** are currently leading an on-going cohort study of outdoor air pollution, including PM_{10} and $PM_{2.5}$, and breast cancer risk in Los Angeles County (Susan G. Komen Foundation, Co-PIs: I. Cheng, A. Wu). This work demonstrates the feasibility and expertise of our team in estimating air pollution exposures and examining their associations with lung cancer incidence.

C1e. Neighborhood characteristics. Our team has a strong interest and history of research in the role of neighborhood social and built environments in relation to cancer incidence and outcomes. Beginning with the development of the California Neighborhoods Data System [93], a collection of small-area level (block group and tract) social and built environment measures for California, **Drs. Gomez, Shariff-Marco, and Cheng** have received funding on over a dozen studies that use these neighborhood data. As evidence of our leadership in the area, we (**Drs. Gomez, DeRouen, and Shariff-Marco**) recently published an invited review in Cancer [85].

C1f. Methodologic work on data pooling and harmonization. Members of the study team have led efforts to pool and harmonize data across diverse sources. **Dr. Gomez** was the site PI, and **Drs. Cheng and Shariff- Marco**, Co-Investigators, on the California Breast Cancer Survivorship Consortium, an effort to pool study interview and cancer registry data on breast cancer survivors

from seven California studies [124-127]. Dr. Gomez, in her capacity as the Greater Bay Area Cancer Registry Co-Investigator, led the Cancer Registry Data Core, and was integrally involved in the development of the protocol for data pooling and harmonization. **Drs. Gomez and Thompson** are investigators in the Oncoshare project, which involved pooling and harmonizing EHR data for breast cancer patients seen at Stanford and PAMF [128]. **Dr. Thompson's** Ph.D. dissertation used data from the Epidemiology of Endometrial Cancer Consortium [129], which included extensive methodological work to identify and evaluate sources of systematic error [130], arising from the pooling and harmonizing of data from disparate sources [131]. **Dr. Cheng** has led harmonization efforts of genetic and epidemiologic data for studies of lung cancer in the Population Architecture Using Genomics and Epidemiology (PAGE) and Transdisciplinary Research in Cancer of Lung (TRICL) consortia. These experiences collectively demonstrate our team's experience in pooling and harmonizing data from diverse sources, recognizing potential biases, and developing plans and methods for addressing biases.

C1g. Research with EHR data. **Dr. Waitzfelder** is a health services research investigator with considerable expertise in population-based studies, particularly involving diabetes and the use of health plan data. She is PI of a pending R01 (score within funding range) that proposes to merge EHR data from Kaiser Hawaii and Sutter Health. **Dr. Thompson** has been working with EHR data from Sutter Health and Stanford University. She has investigated methods for longitudinal analysis of routine data, cohort definition considering in- and outmigration from a health system over time, and handling differences in data quality from combining multiple (seemingly similar) EHR systems (via pooling, harmonization, and linkage). She recently developed an algorithm to classify longitudinal patterns of cancer care into "episodes" using incidence and treatment data from the long term follow-up of 13,000 breast cancer survivors in Oncoshare [132]. She also led an effort to study disparities in cancer screening among disaggregated AANHPI populations, identifying factors associated with screening compliance at the provider- and patient- level using PAMF EHR data [133]. **Dr. Gomez** led a study to link EHR data from Kaiser Permanente Northern California with the California Cancer Registry to assess factors associated with racial/ethnic and nSES differences in receipt of guideline treatment and breast cancer survival [134, 135]. Within Oncoshare, **Dr. Gomez** facilitated the linkage of the Stanford and PAMF EHR data with the California Cancer Registry data [136]. She has also conducted research with the SEER-Medicare database [137], and thus is knowledgeable about the nuances of working with clinical encounter and billing data.

C2. Research Design

C2a. Overview

The objective of our study is to estimate overall and histologic-specific lung cancer incidence rates by smoking status (*Aim 1*) and to identify etiologic risk factors for lung cancer among female AANHPI never smokers (*Aim 2*). We propose to address this objective by pooling longitudinal EHR data from Sutter Health and Kaiser Hawaii. Both systems are part of the HMO Research Network (HMORN), thus facilitating data harmonization through use of the Virtual Data Warehouse (VDW), a tool to facilitate multi-site cancer research through standardized, interoperable data files [138]. Together these databases include 3.2 million adult individuals (*Aim 1*), with nearly 170,000 AANHPI never-smoking females (*Aim 2*), and up to 15 years of follow-up (2000-2015).

In *Aim 1*, by linking to the statewide California Cancer Registry (for Sutter Health) and Hawaii Tumor Registry (for Kaiser Hawaii), we will identify and confirm incident lung cancers among study subjects. Linkage to the statewide cancer registries will provide verified lung cancer

diagnoses, even among those who subsequently dis-enrolled (health system out-migration, e.g., 60% of Kaiser Hawaii members are enrolled for 5+ years). We will use Poisson regression models to estimate cumulative incidence of lung cancer and incidence rates for specific histologic cell-types by sex, race/ethnicity, and smoking status (never, former, current, passive (SHS)). We will also characterize the distribution of sociodemographic, tumor (e.g. stage, histology), and molecular characteristics (e.g. EGFR, ALK mutations) of lung cancer cases by sex, race/ethnicity, and smoking status. Although the focus is on AANHPI females, for comparison and as these data will be readily available as part of this study, we will also conduct analyses for males and females of all racial/ethnic groups.

In *Aim 2*, we will capitalize on the rich clinical data available within the EHRs coupled with neighborhood and environmental data to identify etiologic risk factors for lung cancer among AANHPI female never smokers. We will conduct a longitudinal analysis of lung cancer risk, focusing on six exposure domains: 1) passive smoking; 2) previous lung diseases; 3) infections; 4) reproductive history and hormone exposure; 5) body size; and 6) environmental factors including PM, traffic density, nSES, and ethnic enclave. Using multi-level Cox proportional hazard models, we will estimate the associations of these factors with lung cancer risk. In addition, we will conduct absolute risk modeling to quantify the probability of developing lung cancer for each exposure.

C2b. Description of cohorts

C2b1. Sutter Health medical foundations, Northern California. Sutter Health is a non-profit health system that delivers healthcare coverage in 18 northern California counties (Santa Clara, Santa Cruz, San Mateo, Del Norte, Marin, Sacramento, San Francisco, Alameda, Contra Costa, Merced, San Joaquin, Stanislaus, Amador, Placer, Solano, Sutter, Yolo, and Yuba counties) across 150 ambulatory medical clinics, with more than 3 million patients and over 10 million outpatient visits per year. It comprises five medical foundations: 1) Palo Alto Medical Foundation (PAMF); 2) Sutter East Bay Medical Foundation; 3) Sutter Pacific Medical Foundation; 4) Sutter Gould Medical Foundation; and 5) Sutter Medical Foundation. The demographics of the patient population are generally representative of the underlying population with respect to sex, age and race/ethnicity. The Sutter Health patient population is very diverse, including 19% AANHPIs. Sutter Health patients have insurance plans covered by a payer mix of PPO (42%), HMO (30%), Medicare/Medicaid (23%), self-payers (3%), and other payers (2%). An advantage of the Sutter Health system is that patients are able to remain in the system with their own physician regardless of change in employer-provided health plan. The EpicCare EHR system, designed by the Epic Systems Corporation (Verona, WI), has been in use for nearly 15 years at Sutter Health. It has been active since 2000 at PAMF, and all medical foundations were live on EpicCare as of 2010. For the later adopting foundations, important clinical data (e.g., encounter, procedures, medical history, labs, etc.) from legacy systems (dating back to 1999) were migrated into the new systems, thus EHR data are available for all five foundations starting from at least 2000. The EpicCare system is designed to collect details of all patient encounters, including laboratory results, procedures, medication orders, diagnoses, immunizations, radiologic reports, and routine testing, as well as demographics, medical and surgical history, and transactional detail about care utilization (providers seen, physician notes, dates and times, communications between providers and patients, etc). The PAMFRI Information Management Group maintains access to Sutter-wide EHR data with HIPAA-compliant procedures in place to provide data in limited- and de-identified form to Sutter researchers and their collaborators.

C2b2. Kaiser Hawaii. Kaiser Permanente Hawaii (KPH) was founded in 1958 and is a non-profit, integrated health care delivery system with over 230,000 members, one 285-bed acute care

inpatient facility, 20 outpatient clinics on Oahu, Maui, and Hawaii Island, and numerous independent primary care providers on Kauai, Lanai, and Molokai. About 75% of patient membership lives on the island of Oahu, the remainder on Maui and Hawaii Island. Health plan members are highly representative of the state's population, which is one of the world's most diverse regions (77% minority). In fact, Hawaii's current demographic features reflect some important shifts that are beginning to take place on the mainland, including increasing AANHPI and mixed race/ethnicity populations. In December 2004, Kaiser Permanente began implementing a new integrated EHR system designed by Epic to automate its patient files and make documentation of care more efficient and complete. Epic products have been extensively customized for the needs of KP where it is now referred to as KP HealthConnect (KPHC). Though KPHC was implemented in 2004/2005, KPH has captured inpatient diagnoses and procedures electronically since 1985 and pharmacy, laboratory tests and claims since 1995. KPHC documents all outpatient and inpatient encounters, billing, appointments, patient registration, surgery operating room management and provides the MyChart Web portal for members. The KPH VDW is constructed by extracting data directly from EHRs and reconfiguring it to use standard variable names and coded values Data on health and health care utilization, including hospitalizations, outpatient visits, laboratory tests and values, and pharmacy orders and fills, are included in the VDW.

C2b3. Study Population. Our pooled cohort comprises 1.8 million adult females and 1.4 million adult males, with utilization data from clinical encounters from 2000-2015. Study subjects will be considered eligible if they have had at least one in-person contact with a health care provider with no previous diagnosis of cancer at the time of cohort entry. Among 240,969 AANHPI females (**Table 1**), 214,428 are of single race/ethnicity and 26,541 are of mixed race/ethnicity. Among 177,810 AANHPI males, 155,038 are of single race/ethnicity and 22,772 are of mixed race/ethnicity. Among females, 70% are estimated to be never smokers, 21% as ever smokers (14.8% previous and 6.7% current), and 1.3% exposed to SHS. Among males, 60% are estimated to be never smokers, 29% as ever smokers (18.5% previous and 10.4% current), and 1.7% exposed to SHS.

Table 1. Frequency distributions of study subjects by race/ethnicity, age, and smoking history by sex, Sutter Health and Kaiser Hawaii, 2000-2014

	Females (n)	Males (n)	Total (n)
Total	**1,831,843**	**1,443,361**	**3,275,204**
Cohort			
Sutter Health, Northern California			
Kaiser Hawaii			
Race/ethnicity			
Total AANHPI			
AANHPI			
Asian			
Chinese			
Japanese			
Filipino			
Korean			
Vietnamese			
Native Hawaiian & Pacific Islander			
Native Hawaiian			
Other Pacific Islander			
Other AANHPI*			
AANHPI (mixed race/ethnicity)			
AA & NHPI			
AA & White			
NHPI & White			
Non-Hispanic White			
Non-Hispanic Black			
Hispanic			
Other single race/ethnicity			
Other mixed race/ethnicity			
Unknown/missing			
Age			
<50			
50-59			
60-69			
70-79			
80+			
Smoking history			
Never smoker			
Ever smoker			
Former smoker			
Current smoker			
Passive smoker			
Not Asked			
Unknown/missing			

* "Other AANHPI" includes AANHPI ethnic groups not specifically listed here

For *Aim 1*, we will compute incidence rates by smoking status and race/ethnicity and examine sociodemographic, tumor, and molecular distributions among lung cancer cases. Our study

population will include all racial/ethnic groups and lung cancer cases shown in **Tables 1 and 2**. Males and other races/ethnicities will be examined for comparison. For *Aim 2*, we will conduct a comprehensive analysis of six exposure domains and lung cancer risk among AANHPI never smoking females. Here, we will focus our study population on female AANHPI single and mixed race/ethnicities with sufficient numbers of never smoking lung cancer cases (**Table 3**).

Table 2. Frequency distribution of hospital-reported & projected lung cancer cases* by race/ethnicity and sex, Sutter Health and Kaiser Hawaii, 2000-2015

	Females (n)	Males (n)	Total (n)
Total	**7,180**	**6,315**	**13,495**
Race/ethnicity			
Total AANHPI			
AANHPI (single race/ethnicity)			
Asian Indian			
Chinese			
Japanese			
Filipino			
Korean			
Vietnamese			
Native Hawaiian & Pac Islander			
Native Hawaiian			
Other Pacific Islander			
Other AANHPI*			
AANHPI (mixed race/ethnicity)			
AA, NHPI			
AA, White			
NHPI, White			
Non-Hispanic White			
Non-Hispanic Black			
Hispanic			
Other single race/ethnicity			
Other mixed race/ethnicity			
Unknown/missing			

* Hospital-reported lung cancer cases based on ICD-9 codes; projected through 2015 based on average annual case counts.

** Other AANHPI includes AANHPI ethnic groups not specifically listed

Table 3. Aim 2 projected lung cancer cases among AANHPI female never smokers, by racial/ethnic group, Sutter Health and Kaiser Hawaii, 2000-2015

Race/ethnicity	Counts (n)
Total AANHPI	**488**
AANHPI (single race/ethnicity)	
Chinese	
Japanese	
Filipino	
Native Hawaiian & Pac Islander	
Native Hawaiian	
Other AANHPI	
AANHPI (mixed race/ethnicity)	
AA, NHPI	
AA, White	
NHPI, White	

C2c. Study measures

C2c1. Dependent measure. The primary dependent variable in *Aims 1* and *2* is diagnosis of invasive lung cancer. This will be obtained via linkage to the California Cancer Registry (for Sutter) and the Hawaii Tumor Registry (for Kaiser Hawaii). Population-based cancer reporting is mandated by law in both states, and both registries are a part of the NCI SEER program. From the registries, we will also obtain information on: histology, stage, age and date of diagnosis, sex, race/ethnicity, vital status, and cause of death. For *Aim 1*, we project to have 13,495 total lung cancer cases (**Table 2**). For *Aim 2*, we project to have 488 single and mixed race/ethnicity AANHPI never-smoking (self-reported 'yes/no') female lung cancer cases (**Table 3**).

C2c2. Independent measures. The primary exposures of interest are smoking history, molecular factors, previous or pre-existing lung diseases, infections, reproductive factors and hormone exposure, body size, and environmental factors. Additional primary variables or confounders of interest include sex, race/ethnicity, language preference & translator use, stage, and histology. The source and description of these measures are described in **Table 4**.

Table 4. Source & description of study measures

Study measure	Source, description
Sex	EHR/Virtual Data Warehouse
Race/Ethnicity	EHR/Virtual Data Warehouse, from self-report
Language preference & translator use (as indicator of acculturation)	EHR/Virtual Data Warehouse, from self-report
Smoking history & second-hand smoke exposure Ever smoker Past or current smoker Never smoker Exposure to second-hand smoke	EHR/Virtual Data Warehouse, obtained from patient social history: asked by physician at encounter.

Study measure	Source, description
Molecular factors EGFR ALK	EHR/Virtual Data Warehouse, ICD-9 diagnostic codes from encounters
Infections Pneumonia Tuberculosis Human papillomavirus Human immunodeficiency virus Chlamydia	EHR/Virtual Data Warehouse, ICD-9 diagnostic codes from encounters
Reproductive factors & hormone exposures Estrogen/hormone replacement therapy Parity & number of live births Age at first live birth Age at menarche Average length of menstrual cycle Oral contraceptive use Age at menopause Type of menopause	EHR/Virtual Data Warehouse, diagnostic codes from encounters, pharmacy data, state birth record data (self-report and hospital report), pregnancy information (EHR: CPT procedures for deliveries)
Body size Body mass index	EHR/Virtual Data Warehouse, weight, height, BMI measurements (reported by medical staff)
Particulate matter (California)	California Air Quality Data (California Air Resource Board): measured concentrations of criteria air pollutants from monitoring sites throughout CA.
Traffic count & density (California)	California Dept. of Transportation; Highways and major roadways; Annual average daily traffic, peak hour traffic, and peak month average daily traffic, at street description and post mile level.
Neighborhood SES (California & Hawaii)	2000 Census & 2007-2011 American Community Survey; Census tract-level composite measure derived from principal components analysis and comprising income, education, poverty, employment, occupation, housing and rent values.
Ethnic enclave (California & Hawaii)	2000 Census & 2007-2011 American Community Survey; Census tract-level composite measure derived from principal components analysis, California-specific index comprises 4 components including AANHPIs, recent immigrants, AANHPIs who speak limited English, and AANHPI households that are linguistically isolated; a separate index for Hawaii will be derived.

Regionally-distributed pollutants. Monthly and yearly ambient data on PM_{10} and $PM_{2.5}$ collected from air monitoring stations will be used to create estimates of exposure to regionally-distributed pollutants using kriging interpolation. These monthly averaged concentrations will be spatially interpolated between stations using an empirical Bayesian kriging model implemented in ArcGIS (ESRI, Redlands, CA) to account for the uncertainty of semivariogram estimation [139]. Monthly or

yearly-specific exposure averages will be based on the duration of residential history from enrollment to address change, end of follow-up, or address at disease diagnosis. We have experience applying this spacetime interpolation technique [140], which has been shown to increase mean precision compared to ordinary kriging [141]. *Geocoding*. All addresses, including changes in addresses for individuals over time (maintained for billing purposes), will be geocoded to a coordinate (latitude & longitude) and assigned to a Census 2000 (for addresses from 2000-2005) or 2010 (for addresses from 2006+) tract and block group. Kaiser Hawaii addresses are routinely geocoded via ESRI and updated for each individual. For the Sutter Health addresses, we will use the Texas A & M batch geocoder (http://geoservices.tamu.edu/Services/Geocode/), used by most state cancer registries, including California and Hawaii. We will manually geocode those addresses that do not successfully batch geocode.

C2d. Statistical analysis

All analyses will be done in SAS 9.3 or R. Our biostatistician, **Dr. David Nelson**, is experienced in working with longitudinal and geospatial data. Through modeling methods developed by Co-Investigator **Dr. Caroline Thompson**, we will evaluate the influence of suspected biases (confounding, selection bias) that may vary by data source, either as a result of data harmonization decisions or underlying data generating mechanisms [130, 131]. We will use incidence density Poisson regression analysis for calculating incidence rates; descriptive analyses and logistic, linear, or polytomous regression for examining distributions of socio-demographic and clinical characteristics; and Cox proportional hazards regression for examining etiologic risk factors. We will also conduct sensitivity analyses to evaluate the impact of missing data on smoking status.

C2d1. Data extraction, pooling, and harmonization. As both Sutter and Kaiser Hawaii utilize the VDW [138], we will use the distributed data analysis method to develop and share SAS code for extracting the data files from each site, thereby ensuring that the variables are abstracted and coded consistently across the two sites. We will extract, pool, and harmonize data, as well as conduct descriptive analyses to understand patterns of missing data. If patterns of missing data are systematic by geographic location, we will consider the underlying data generating mechanisms attributable to these patterns and account for them in sensitivity analyses of our multivariable risk models (see section *C2d4*).

C2d2. Aim 1 analyses. The goal of *Aim 1* is to estimate cumulative incidence and incidence rates over time for pre-defined sub-cohorts of interest and pre-specified calendar and follow-up time periods. We will assume complete follow-up by the cancer registries, thus participants will be followed from time of cohort entry (at entry into the health plan or first visit to the healthcare facility) until the date of lung cancer diagnosis or censoring on the date of death from another cause. The main analytic approach will be either Poisson regression or negative-binomial regression, depending on whether overdispersion is detected [142]. Individual cohort data will first be reduced to a sufficient dataset, in which each observation will describe a stratum and consist of the number of events and of the person-years of risk, accompanied by a set of covariates that define the characteristics of the stratum. Minimally, these covariates will include race/ethnicity, age group, and year of cohort entry, and geographic location (CA or HI). In addition, stratified analysis will be conducted by sex, race/ethnicity, and smoking status. Any required two dimensional smoothing of rates by age group and year will be performed using the smoothing spline features in the R packages *mgcv* [143], or *mortalitySmooth* [144]. We will also examine distributions of socio-demographic, tumor, and molecular characteristics among lung cancer

cases by race/ethnicity, sex, and smoking status, and test for statistically significant differences using chisquared statistics.

C2d3. Aim 2 analyses. The goal of *Aim 2* is to understand the etiology of lung cancer incidence among neversmoking AANHPI females, especially with respect to passive smoking, previous lung diseases, infections, reproductive history, body size, and contextual factors such as exposure to particulate matter, traffic density, nSES and ethnic enclave status. Multivariable and multi-level Cox proportional hazards regression models will be used to regress lung cancer events on predictors of interest, while adjusting for important covariates, and generate hazard rate ratios (HR) and 95% confidence intervals. All analyses will control for or be stratified on sex, race/ethnicity, and/or histologic cell type. Proportionality in hazards will be checked using Schoenfeld residuals for all independent variables. We will apply corrections to control the False Discovery Rate for multiple testing [145]. In addition, we will construct models for specific AANHPI female ethnic groups and test for heterogeneity in associations by AANHPI ethnicities as power allows.

Because we are evaluating time-dependent individual- and contextual-level predictors, study subjects will additionally be eligible for censoring upon health system out-migration, or death, whichever comes first. Health system out-migration will be determined algorithmically as 1) the date of disenrollment from Kaiser HI, or 2) after an extended period of inactivity from Sutter. The time scale will be age (in days) from date of cohort entry (at entry into the health plan or first visit to the healthcare facility) until the end of follow-up, and adjusted for hypothesized predictors, key covariates, and health system. Statistically significant effects will be detected by comparing sequences of nested models using (partial) likelihood ratio tests. Trends for continuous variables will be evaluated by representing the variable as an orthogonal polynomial of modest degree and testing the significance of individual coefficients.

Variable selection. Univariable analyses for variables of interest will be performed. Scatterplot matrices and rank correlations will be used to assess relationships among exposure variables. In addition, variable reduction methods like the lasso and Random Forests will be used to prune the large number of covariates by eliminating spurious predictors from models and by grouping highly correlated predictors [146, 147].

Analyzing continuous exposures. During initial analyses, graphical methods and semi-parametric models, such as smoothing splines, will be used to assess the gross structure of exposure-outcome relationships. Based on the results of these graphical analyses, continuous variables (e.g., ambient air exposures) will be analyzed in two distinct ways. First, in order to assess potential threshold effects, continuous variables will be modeled as two-level factors with varying proportions of exposed subjects. Second, in order to assess linear doseresponse effects (on a log scale), they will be log-transformed to approximate normality. After log transforming, they will be modeled by orthogonal polynomials of degree 1 or 2.

Multi-level analyses. The joint effects of individual risk factors and environmental measures (PM, traffic density, nSES, ethnic enclave) will be modeled by multi-level models (i.e., frailty or random effects models), which will allow the concurrent estimation of effects of factors at multiple, nested, levels of data organization. Because some of the area-level measures are based on arbitrary geographical spatial definitions, there is a potential for residual spatial autocorrelation (RSA), or the propensity for values in neighboring areas to be similar, which may affect assumptions about the independence of residuals. If significant RSA is detected, the final model will be refit, using frailty models and robust variance estimates that account for the correlation.

Absolute risk modeling. Because our dataset will be a prospectively modeled cohort, we will broaden the class of models to allow the direct modeling of absolute risks and their dependence on covariates. Compared to Cox regression, these models have the benefit of isolating the effects of modifiers of excess risk due to exposures from modifiers of baseline rates. We can use these models to test, for example, whether PM affects baseline risk, the passive smoking effect, or both. We can also readily provide summaries of the total excess number of cases due to an exposure in the population with a certain age and risk factor structure.

We will also calculate the population attributable fraction (PAF) of select individual or groups of risk factors in the total population of AANHPI female never smokers. The PAF will provide information regarding the excess incidence of lung cancer attributable to each risk factor as a percentage of lung cancer incidence in the total population of AANHPI female never smokers (i.e. the reduction of lung cancer incidence that would be achieved in the population if the risk factor in question were eliminated, assuming the selected risk factor is causally associated with lung cancer incidence). Given that lung cancer is a rare event, the PAF can be calculated simply as the difference in the hazard of lung cancer among the total population and the unexposed population divided by hazard in the total population, so that $\mathrm{PAF}=pd\left(\frac{HR-1}{HR}\right)$; where pd = proportion of cases exposed to risk factor [148-150]. For multiple category exposures, we can also calculate $\mathrm{PAF} = 1-\sum_{i=0}^{k}\frac{pd_i}{HR_i}$; where pd_i = proportion of cases falling into *i*th exposure level; HR comparing *i*th exposure level with unexposed group (*i*=0) [150]. In the event it is reasonable to expect that the PAF will change during follow-up time, we will calculate the PAF as a function of time using components of the Cox model available in SAS [148, 151, 152]. We will also explore more general PAF functions using the methods described by Chen et al. [153] and implemented in R.

C2d4. Sensitivity analyses, missing data. EHR data are distinct from data collected for research purposes because they are generated as a result of healthcare utilization. In EHR data, it is not always clear whether a data point is missing because the subject is unexposed, or because the detail was never disclosed [154]. Thus, "missing data" will be classified on a spectrum and handled accordingly. Customary baseline covariates (age, sex) will be assumed to be missing at random and will be addressed using multiple imputation via chained equations in the R package MICE. Missing race/ethnicity will be imputed using an algorithm that incorporates surname, use of interpreter and languages spoken. Study subjects missing data on clinical predictors such as prior infections, lung diseases, etc. will be considered to be unexposed as long as they are still enrolled in the health system at the time of assessment. To test our assumptions about assigning study subjects with absent clinical data as unexposed, we will perform sensitivity analysis using a measure of frequency of healthcare utilization as an interaction term in models where substantial absent data have been assigned as unexposed. Because smoking data are likely to be more complete for lung cancer cases than non-cases, smoking status will be considered missing not at random and handled by sensitivity analyses for unmeasured confounding to determine the robustness of complete case models to those adjusted for differential missing smoking data by lung cancer status. We will conduct sensitivity analyses using established methods of external adjustment [155] as well as more advanced novel methods of data augmentation, developed by Dr. Thompson [131].

C2e. Minimum detectable effect

As *Aim 1* is a descriptive aim to estimate incidence rates for subgroups of interest, there is no "effect" in play and no minimum detectable effect to estimate. For *Aim 2*, all power calculations are based on the algorithms described in Therneau and Grambsch [156], and are expressed as a Minimum Detectable Hazard Ratio (MDHR) at 80% statistical power and an overall test size of 5% for given two-level exposure prevalences. In this context, the MDHR is a function of (1) expected number of events, and (2) the exposure prevalence. The expected number of events is conservatively based on expected follow-up. Overall, for all never-smoking female AANHPIs (projected lung cancer cases=488; **Table 3**), we estimate MDHRs of 1.3 to 1.5 for exposures that range in prevalence from 30% to 10%. For the four AANHPI ethnic groups with the largest expected case counts (Chinese, Japanese, Filipina, NHPI), we estimate MDHRs=1.9-2.2 and MDHRs=2.6-3.2 for exposure prevalence of 30% and 10%, respectively. These MDHRs are in line with reported associations such as the HR=2.46 between PM10 and lung cancer risk among Japanese Americans in the Multiethnic Cohort (n=30 never-smoking Japanese American female lung cancer cases; Section *A5f*). Thus, affirming our study will have acceptable study power.

D. Expected Outcomes

In the short-term, through our assembly of the largest epidemiologic resource to date of AANHPI lung cancer cases and population counts, with essential information on smoking history, this study will fill a critical gap in understanding the burden of lung cancer among AANHPIs—the fastest growing racial/ethnic group in the U.S. This study is timely given that AANHPIs are estimated to grow over four times as rapidly as the total U.S. population with increasing trends in lung cancer incidence for certain female AANHPI ethnic groups. Our study is also expected to shed light on the underlying causes of lung cancer among never-smoking AANHPI females. Such findings may have direct translational impact on informing opportunities for cancer control and prevention through modifying health behaviors and the surrounding environment, and potentially informing guidelines for lung cancer screening. In the long-term, the infrastructure we will develop for a racially/ethnically diverse cohort of over 3 million health plan members, merging clinical EHR, molecular, neighborhood, and cancer surveillance data, will serve as a unique resource to study racial/ethnic disparities in cancer risk and survival such as extending this study into the evaluation of molecular tumor characteristics and prognostic factors. Also, with follow-up of these subjects for various health outcomes, the value of this resource will further increase to facilitate additional research inquiries for future studies.

E. Potential Problems & Alternative Strategies

We recognize the "representativeness" of EHR data is of relevance in describing disease occurrence and making inference about a larger source population. To assess representativeness of our EHR-based study, we will make comparisons to statewide neighborhood contextual data and to cancer registry data for lung cancer patients. We have recently used similar methods to show in a manuscript (under review) that breast cancer patients in the northern California Kaiser medical system are representative in socio-demographic and clinical characteristics to the general population of cancer patients as described in the regional cancer registry. Furthermore, recent debates in the epidemiologic literature as to the importance of the generalizability of association analyses of cohort studies have suggested that study populations that are so heterogeneous that they are "representative" can hinder internal validity due to lack of sample size in one or more subgroups of interest or difficulties with confounding control [157]. Despite our attempts to maximize sample sizes in the

targeted ethnicities, we acknowledge the problem of small numbers and limited power to study certain AANHPI groups. This is a recognized challenge by the NIH with its responsibility to conduct research to improve the health of all, not solely the health of the majority or those who are easy to identify [88]. In a recent editorial, Srinivasan et al. [88] discussed the importance of small data of subpopulations and encouraged research in these groups, using innovative data analytic approaches such as integrative data analysis [158], where independent data sets are combined to maximize sample sizes, to improve precisions and internal validity, such as we have proposed here. In the planning for this proposal, we had evaluated several additional EHR and cohort data sources, but those were not included primarily due to high proportions of missing data for race/ethnicity and/or smoking status (such as the California Kaiser systems), small numbers of AANHPIs, and/or lack of novel exposures for evaluation. The inclusion of data from another large health system in the same geographic area would also present issues of study subject overlap. Another problem that could arise may be that baseline individual-level and neighborhood-level exposures may not be representative of exposures over time. Therefore, we would repeat this analysis incorporating time-dependent measures of exposures. Additional threats to internal validity include unmeasured confounding/missing data, which will be handled using sensitivity analyses as described in section *C2d4*.

Despite these limitations, this efficient study nonetheless will represent to-date the largest resource to document lung cancer incidence by sex and smoking status among AANHPI ethnicities, and to allow us to study the relative contributions of risk factors for lung cancer among never-smoking AANHPI females.

F. Timeline

Activity (Top row: Year; Bottom row: Month)	**1**		**2**		**3**	
	6	**12**	**18**	**24**	**30**	**36**
IRBs, data use agreements, develop common data dictionary	**X**					
Extract EHR/VDW data, link to state cancer registries; geocode and derive environmental variables; pool and harmonize data		**X**	**X**			
Conduct analysis for Aim 1			**X**	**X**		
Report results for Aim 1; conduct analysis for Aim 2				**X**	**X**	
Report results for Aim 2					**X**	**X**

HUMAN SUBJECTS

Introduction: The proposed study uses quantitative methods to analyze data from three different sources: cancer registry linked EHR data from Northern California Sutter Health and Kaiser Permanente Hawaii, and contextual level data collected to characterize study subject's neighborhoods and environment. The population will include an expected total of 3 million individuals; approximately 13,500 lung cancer cases will be identified. All of the underlying sources of data already exist, and contain protected health information (PHI), however, Palo Alto Medical Foundation Research Institute (PAMFRI) and Kaiser Permanente Hawaii have developed HIPAA-compliant procedures to link files and prepare de-identified datasets for research use, these include removal of PHI and obfuscating dates. In order to ensure the environmental and neighborhood contextual attribute data are linked appropriately, patient address history (including latitudes and longitudes) will be shared with trained staff members at Cancer Prevention Institute of California (CPIC), however small-area geographical identifiers will be removed prior to analyses of these data by researchers at CPIC, PAMRI and/or Kaiser Hawaii. There will be no human subjects contact for this study. Human subject involvement is limited to the use of their EHR data with linkage to neighborhood environment data and cancer registry data. Information from electronic health records will include demographic factors, health behaviors, medical conditions, and tumor molecular characteristics. The racial/ethnic distribution of study participants is provided in the Targeted Enrollment table.

Potential risks to the subjects: We expect there to be minimal risk to the subject, as it involves only analysis of existing data. No patients will be recruited; no interventions are planned. In regards to potential for loss of confidentiality, we have extensive experience in maintaining confidentiality of epidemiological and molecular data and believe that the risk is very low. We will take various steps to minimize the possibility of loss of confidentiality, as discussed in the next section.

ADEQUACY OF PROTECTION AGAINST RISKS

Informed consent: For the purpose of this study, there will be no extra contact with the study subjects. Approval to involve human subjects in this research study will be obtained from CPIC, Sutter Health, Kaiser Permanente Hawaii, and the California Protection for Human Subjects prior to data pooling and analyses. IRB approval, including waivers of informed consent and HIPAA authorization for this specific proposal will be obtained prior to initiating the proposed study.

Protection against risk: Pooling and linking of Sutter Health and Kaiser Permanente Hawaii data will be conducted at CPIC and no identifying information of Sutter Health participants will be released to Kaiser Permanente Hawaii; similarly, no identifying information of Kaiser Permanente Hawaii participants will be released to Sutter Health. Computerized databases containing identifying information will be maintained on the CPIC network and will be accessible only to eligible study staff via encrypted, password-secured computers. The databases themselves will also be password-secured. Analysis datasets will be generated using deidentified data files. Study staff will be trained to follow established CPIC, Sutter Health, and Kaiser Permanente Hawaii confidentiality procedures and will sign confidentiality agreements. No preliminary or final results will be released or published with identifying information and all epidemiologic data will be presented in aggregate form. All research staff will complete human subjects training.

Discussion of why risks are reasonable in relation to anticipated benefits to subjects: While the study participants themselves are unlikely to benefit directly from the study, the

potential benefits are nevertheless numerous. We have assembled data form a large group of both male and female participants of several major racial/ethnic groups in California and Hawaii. This study provides a unique opportunity to quantify the burden of lung cancer among Asian American, Native Hawaiians, and Pacific Islanders and investigate the epidemiology of lung cancer among never smoking females of these specific racial/ethnic groups. The study assembled here is particularly well suited for this type of epidemiologic analysis of lifestyle and environmental factors and characterization of molecular markers. In view of the minimal risks to the participants and the substantial potential for the study to yield important and useful data to the scientific community for the ultimate prevention of this common deadly cancer, the benefits from the proposed study should greatly outweigh the risks.

Importance of the knowledge to be gained: Large studies of well-characterized subjects, such as the one proposed here, should help elucidate the cause of lung cancer. With a better understanding of these causes, we should be able to develop more effective diagnosis, prevention and treatment strategies.

Data and Safety Monitoring Plan
Not applicable.

INCLUSION OF WOMEN
Women are included in this proposal; there was no preferential inclusion by gender. The gender distribution is provided in the table below (see Targeted Enrollment Table).

INCLUSION OF MINORITIES
This study includes Asian Americans (Asian Indian, Chinese, Japanese, Filipino, Korean, and Vietnamese), Native Hawaiians, other Pacific Islanders, Non-Hispanic Blacks, and Hispanics from Hawaii and California. The distribution by ethnicity and race for these subjects is provided in the table below (see Targeted Enrollment Table).

OMB Number: 0925-0002

Planned Enrollment Report

Study Title: Lung cancer in never smokers: incidence, risk factors, and molecular characteristics in Asian American, Native Hawaiian and Pacific Islander females

Domestic/Foreign: Domestic

Comments: Our planned enrollment also includes 622,909 females and 565,495 males coded in the electronic health records as having other, unknown, or missing race/ethnicity. This brings our total planned enrollment to 3,275,204.

Racial Categories	Ethnic Categories				Total
	Not Hispanic or Latino		Hispanic or Latino		
	Female	Male	Female	Male	
American Indian/Alaska Native	0	0	0	0	**0**
Asian	190084	131335	0	0	**321419**
Native Hawaiian or Other Pacific Islander	24344	23703	0	0	**48047**
Black or African American	50482	31977	0	0	**82459**
White	690388	492443	154032	106560	**1443423**
More than One Race	99604	78244	0	0	**177848**
Total	**1054902**	**757702**	**154032**	**106560**	**2073196**

Study 1 of 1

INCLUSION OF CHILDREN

Children were excluded from this proposal because the occurrence of lung cancer is rare in children.

CONSORTIUM/CONTRACTUAL ARRANGEMENTS

This project involves subcontracts to the Palo Alto Medical Foundation Research Institute (PAMFRI), Kaiser Hawaii, and Stanford University. The role of these subcontractors in this study is described in detail in this application and in the budget justification.

Kaiser Hawaii will share limited data from their electronic health records (EHR) with CPIC investigators for the purpose of addressing the project's Specific Aims. Specifically, this will involve implementing data use agreements and facilitating IRB applications; working with collaborators at CPIC and PAMFRI to develop common SAS code for extracting the relevant data items from the Virtual Data Warehouse (VDW); accessing and providing updated cancer registry data; and providing guidance on data pooling and harmonization. Dr. Waitzfelder will also participate as Co-Investigator throughout the study by participating in regular study team meetings, reviewing and interpreting results, and co-authoring manuscripts.

PAMFRI will share limited data from their EHR with CPIC investigators for the purpose of addressing the project's Specific Aims. Specifically, this will involve implementing data use agreements and facilitating IRB applications; working with collaborators at CPIC and Kaiser Hawaii to develop common SAS code for extracting the relevant data items from the VDW; accessing and providing updated cancer registry data; and providing guidance on data pooling and harmonization. Drs. Thompson and Luft will also participate as Co-Investigators throughout the study by participating in regular study team meetings, reviewing and interpreting results, and co-authoring manuscripts.

Co-Investigators Drs. Wakelee, Patel, and Haile at Stanford University will provide critical content expertise and input by participating in regular study team meetings, reviewing and interpreting results, and co-authoring manuscripts.