Title

Integrating Electronic Health Record, Cancer Registry, and Geospatial Data to Study Lung Cancer in Asian American, Native Hawaiian, and Pacific Islander Ethnic Groups

Permalink
https://escholarship.org/uc/item/11h9z1pk

Journal
Cancer Epidemiology Biomarkers & Prevention, 30(8)

ISSN
1055-9965

Authors
DeRouen, Mindy C
Thompson, Caroline A
Canchola, Alison J
et al.

Publication Date
2021-08-01

DOI
10.1158/1055-9965.epi-21-0019

Peer reviewed

Published in final edited form as:

Cancer Epidemiol Biomarkers Prev. 2021 August ; 30(8): 1506–1516. doi:10.1158/1055-9965.EPI-21-0019.

Integrating electronic health record, cancer registry, and geospatial data to study lung cancer in Asian American, Native Hawaiian and Pacific Islander ethnic groups

Mindy C. DeRouen[1,2], **Caroline A. Thompson**[3,4,5], **Alison J. Canchola**[1,6], **Anqi Jin**[5], **Sixiang Nie**[7], **Carmen Wong**[7], **Jennifer Jain**[1], **Daphne Y. Lichtensztajn**[1,6], **Yuqing Li**[1], **Laura Allen**[1], **Manali I. Patel**[8,9], **Yihe G. Daida**[7], **Harold S. Luft**[5], **Salma Shariff-Marco**[1,2,6], **Peggy Reynolds**[1,2], **Heather A. Wakelee**[8], **Su-Ying Liang**[5], **Beth E. Waitzfelder**[7], **Iona Cheng**[*,1,2,6], **Scarlett L. Gomez**[*,1,2,6]

[1]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA, USA

[2]Helen Diller Family Comprehensive Cancer Center, University of California San Francisco, San Francisco, CA, USA

[3]San Diego State University School of Public Health, San Diego, CA, USA

[4]University of California San Diego School of Medicine, San Diego, CA, USA

[5]Sutter Health Palo Alto Medical Foundation Research Institute, Palo Alto, CA

[6]Greater Bay Area Cancer Registry, University of California San Francisco, CA, USA

[7]Kaiser Permanente Hawai'i Center for Integrated Health Care Research, Honolulu, HI, USA

[8]Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA

[9]VA Palo Alto Health Care System, Palo Alto, CA, USA

Abstract

Background: A relatively high proportion of Asian American, Native Hawaiian, and Pacific Islander (AANHPI) females with lung cancer have never smoked. We used an integrative data approach to assemble a large-scale cohort to study lung cancer risk among AANHPI by smoking status with attention to representation of specific AANHPI ethnic groups.

Methods: We leveraged electronic health records (EHRs) from two healthcare systems—Sutter Health in northern California and Kaiser Permanente Hawai'i— that have high representation of AANHPI populations. We linked EHR data on lung cancer risk factors (i.e., smoking, lung diseases, infections, reproductive factors, and body size) to data on incident lung cancer diagnoses from statewide population-based cancer registries of California and Hawai'i for the period 2000–

Corresponding author: Mindy C. DeRouen; 550 16th Street, San Francisco, CA 94158; 415-935-4787; mindy.hebert-derouen@ucsf.edu.

[*]Indicates equal contribution to senior authorship

Conflict of interest statement: The authors declare no potential conflict of interest.

2013. Geocoded address data were linked to data on neighborhood contextual factors and regional air pollutants.

Results: The dataset comprises over 2.2 million adult females and males of any race/ethnicity. Over 250,000 are AANHPI females (19.6% of the female study population). Smoking status is available for over 95% of individuals. The dataset includes 7,274 lung cancer cases, including 613 cases among AANHPI females. Prevalence of never-smoking status varied greatly among AANHPI females with incident lung cancer, from 85.7% among Asian Indian to 14.4% among Native Hawaiian females.

Conclusion: We have developed a large, multilevel dataset particularly well-suited to conduct prospective studies of lung cancer risk among AANHPI females who never smoked.

Impact: The integrative data approach is an effective way to conduct cancer research assessing multilevel factors on cancer outcomes among small populations.

Keywords

Electronic health record cohort; Asian American; Native Hawaiian; and Pacific Islanders; lung cancer in never smokers; integrated data analysis; multilevel

Asian American, Native Hawaiian, and Pacific Islander (AANHPI) females who never smoked may experience a particularly high burden of lung cancer: while an estimated 20% of all females with lung cancer in the United States (U.S.). have never smoked, previous reports have suggested that up to 70% of AANHPI females with lung cancer have never smoked (1,2). Putative risk factors have been identified for lung cancer among those who have never smoked (i.e., passive smoking, air pollution, radon, family history of lung cancer, and cooking oil fumes) (3–5). However, the degree to which these and other suspected risk factors contribute to lung cancer among those of specific AANHPI ethnic groups who have never smoked remains largely unknown, with much of our knowledge based on studies in Asia (3–5). Moreover, in the U.S., AANHPIs are individuals with considerable diversity in country of origin, socioeconomic levels, cultural beliefs and behaviors, degree of English proficiency, immigration experience, generational status, and acculturation (6–9). Proportions of AANHPIs that identify as multiple races are uniquely expanding, with increases of 60% among Asian Americans and 44% among Native Hawaiians and Pacific Islanders from 2000 through 2010 (10,11). Despite this substantial diversity, research has mostly considered AANHPIs as one aggregate group, which masks health inequities that exist across ethnicities. Disaggregation of this population in recent studies shows substantial health and exposure disparities (7,8,12) that reinforce the importance of considering specific AANHPI ethnicities in research (13,14).

The study of lung cancer among AANHPI females who have never smoked has been hindered by the lack of a sufficiently-large and representative data source to document population-level incidence rates of lung cancer stratified by sex, detailed race/ethnicity, and smoking status (10,15). Integrative data analysis, which combines data from multiple sources to enrich the number of observations or data on explanatory variables, takes advantage of existing resources (e.g., health care data) to examine rare diseases among “small populations” (16,17) of interest, like AANHPI ethnic groups (16–22).

We assembled a large-scale, multilevel electronic health record (EHR)-based cohort to facilitate research on the incidence and etiology of lung cancer among specific AANHPI ethnic groups. The cohort comprises males and females of any race/ethnicity but is designed specifically to quantify the burden of lung cancer among detailed single- and multi-racial/ethnic AANHPI groups. The dataset contains (a) EHR data from two large healthcare systems linked to (b) their respective statewide cancer registries and (c) geospatial data. The healthcare systems—Sutter Health in northern California and Kaiser Permanente Hawai'i—were specifically selected for their robust AANHPI representation and categorization, mature EHRs, and research infrastructure. We describe our process of data integration with an emphasis on methodology and the generalizability of the resulting pooled cohort, and present proportions of persons with lung cancer who never smoked for groups according to sex and detailed race/ethnicity.

METHODS

All aspects of the study protocol were approved by the IRBs of the State of California Protection for Human Subjects; University of California, San Francisco; Sutter Health; Kaiser Permanente Hawai'i; and the Hawai'i Medical Association.

Cohort population

The cohort was assembled with data for Sutter Health and Kaiser Permanente Hawai'i patients with an in-person encounter any time from 2000 through 2013. Sutter Health is a not-for-profit multispecialty healthcare delivery system in northern California serving 23 northern California counties through five geographically-based medical foundations; Sutter patients make up 36% of the catchment area population (23). Kaiser Permanente Hawai'i is a not-for-profit, integrated health care delivery system with over 254,000 members and covers 18% of the underlying catchment population. Cohort inclusion and exclusion criteria are detailed in Figure 1. The final pooled cohort includes N=2,211,476 (by system: Sutter Health n=1,871,175, Kaiser Permanente Hawai'i n=340,301). The date of the first in-person encounter was considered the date of cohort entry (baseline date), with follow-up through December 31, 2013 or date of death. Median follow-up was 4.8 years with a maximum of 14 years.

Cancer registry linkages and incident lung cancer diagnoses

All Sutter Health patients were previously linked to the California Cancer Registry for all invasive cancers diagnosed 1988–2013 (23). The Kaiser Permanente Hawai'i cohort was linked to the Hawai'i Tumor Registry for lung cancers diagnosed 1973–2013. For all lung cancer cases, cancer registry data included date of diagnosis, tumor stage (localized, regional, remote), and tumor histology (small cell lung carcinoma, adenocarcinoma, squamous cell carcinoma, large cell or other cell carcinoma, and unspecified) (24). Incident lung cancer (n=7,274) was defined as a diagnosis of invasive lung or bronchus carcinoma (International Classification of Disease for Oncology, third edition, site codes C34.0–34.9; excluding histologic codes: 8580–9999 and 8500) occurring during the study period.

EHR Data Elements

Both Sutter Health and Kaiser Permanente Hawai’i use EHR systems designed by Epic (Epic Systems Corporation, Verona, WI). Sutter Health’s EpicCare EHR system was in three of five Sutter Health foundations by 2000 and in all by 2008. Kaiser Permanente Hawai’i EHR, KP HealthConnect (KPHC), was fully implemented in 2004; however, inpatient diagnoses and procedures have been captured electronically since 1985 and pharmacy, laboratory tests, and claims since 1995.

Individual race/ethnicity:

For both Sutter Health and Kaiser Permanente Hawai’i, patients’ race and ethnicity are collected through self-report either in person or online and allow patients to indicate more than one race; only the most recently reported data is retained. We developed a race/ethnicity classification algorithm (Figure 2) that prioritizes small AANHPI populations and distinguishes between single- and multi-racial/ethnic groups. Individuals indicating *any* Pacific Islander group (alone or in combination with other racial groups) were classified as Native Hawaiians and Pacific Islanders (NHPI) and were further disaggregated into categories of *any* Native Hawaiian (NH) or *any* Pacific Islander (PI) not including NH. Individuals who were not NHPI, but indicating *any* Asian American group were categorized as Asian American, even if they also identified with another race group. Among Asian Americans, individuals who indicated a single Asian group were classified as solely: Asian Indian, Chinese, Filipinx, Japanese, Korean, or Vietnamese, or Other single Asian group. Individuals indicating multiple races that included an Asian group were categorized as either multiple Asian groups only (Asian, multiple group) or as Asian and non-Asian groups (Asian and non-Asian multiple). Individuals not categorized as AANHPI were categorized as non-Hispanic White, Black, Hispanic, multiple (non-AANHPI) races, Other (including American Indian and Alaska Native), and Unknown/missing. With this algorithm, the AANHPI and Black categories included Hispanics and non-Hispanics; the Hispanic category included individuals who indicated Hispanic ethnicity and White race or ‘Other’ race (including American Indian and Alaska Native) as well as those with unknown/missing race.

Individual smoking status:

Patient smoking status is routinely collected and recorded in the social history module of the EHR, with mutually exclusive categories of current, former, passive, never, unknown/missing, and not asked. The EHR retains all answers and corresponding collection dates. For each individual, we extracted up to two non-missing smoking status values (current, former, passive, never): (1) the first available value recorded on the day of or after the baseline visit and (2) the last available recorded before date of lung cancer diagnosis, death, or study end (December 31, 2013). For individuals with discordant entries at these two time points, we applied a simple algorithm to define a single smoking status value *(ever, never*) (Supplemental Table 1a–1c). Those with “current” or “former” for either smoking status value were categorized as *ever* smoking. Those with “never” or “passive” for both status values, or with “never” or “passive” for one value and a non-informative value for the other were categorized as *never* smoking. Because “passive” smoking was only available as a mutually exclusive category of smoking and only 1.0% of the cohort had a status of “passive” smoking, we did not utilize this categorization further. N=2,118,945 (95.5%) individuals in the pooled cohort were assigned one of these smoking status categories, the remainder were coded as “unknown” (Supplemental Table 1a).

Additional EHR elements:

At both sites, self-reported language preference is collected via free text and patients report whether a translator is requested (no/yes) (25); only the most recently reported version of this information is retained. We derived a 3-category ordinal variable to serve as a proxy for individual acculturation (English preference, non-English preference/no translator request, and non-English preference/translator request) (23,26).

To study lung cancer risk factors specifically among AANHPI females who never smoked, additional EHR data were extracted (Figure 1). Data on medical history, diagnostic codes and medication orders recorded from the baseline date to the end of follow-up (December 31, 2013) were extracted to classify exposure to previous lung diseases (chronic obstructive pulmonary disease, emphysema, chronic bronchitis, asthma, and idiopathic pulmonary fibrosis), infections (pneumonia, tuberculosis, chlamydia, human papilloma virus (HPV) and human immunodeficiency virus (HIV)), reproductive factors, and body mass index. Details of these disease definitions are presented in Supplementary Table 2.

Geocoding and geospatial data:

Geocoding: For the Sutter Health cohort, individuals with a California residential address available at the baseline visit were geocoded with baseline latitude/longitude coordinates. For N=1,873,650 total in-state addresses (representing addresses for Sutter patients before cohort exclusions were applied), batch geocoding was performed with ArcGIS (27) to a point address (n=1,269,578) or street locators (n=436,487). Manual geocoding resulted in another n=20,216 addresses matched to point address or street locators. A total of 107,248 (5.7%) were determined to be not geocodable (e.g., post-office boxes). This resulted in a total of n=1,726,281 successfully geocoded addresses (97.7% of geocodable addresses). Baseline addresses from 2000–2005 were assigned to Census 2000 tracts and those from 2006–2013 were assigned to Census 2010 tracts. After cohort exclusions were also applied (Figure 1), there were n=1,721,000 individuals from Sutter Health assigned to a baseline census tract of residence.

Kaiser Permanente Hawai’i patient addresses are geocoded centrally using ArcGIS; census tract identifiers for baseline addresses were provided for the pooled cohort dataset. After excluding n=3,228 with addresses outside of Hawai’i, n=35,344 with missing addresses, and cohort exclusions; there were n=301,729 individuals in the Kaiser Permanente Hawai’i cohort with a 2010 Census tract assigned in Hawai’i. The final pooled cohort comprised n=2,022,729 individuals assigned to a census tract.

Mapping: We created census tract-level maps of the geographic distribution of the cohort with individuals’ baseline addresses. For each state/site, ArcGIS software (27) was used to map frequency quintiles for the total cohort population and AANHPI female cohort population. To display a single set of maps, 2010 Census tract boundaries were used for all individuals.

Area-based measures: Individuals' baseline geocoded addresses (addresses available at the time of their first in-person encounter with the healthcare institution during the study period) were linked via census tracts to area-based measures (neighborhood socioeconomic status [nSES], ethnic enclave, ethnic composition, traffic density) (28–30). Neighborhood SES, a composite index developed previously from principal components analysis, incorporates area-level information on education, occupation, employment, household income, poverty, housing value, and rental value from the Census 2000 Summary Files, and American Community Survey (ACS) 2007–2011 data (28,29). Each census tract is assigned the appropriate composite score for nSES and then categorized into quintiles based on the statewide distribution of scores across all census tracts separately for each decennial year (2000,2010). Ethnic enclave is a composite measure of proportions of Asian, recent immigrants and households with limited English or that are linguistically isolated and speak an API language. Statewide quintiles are used to identify ethnically distinct enclaves (quintiles 4 and 5). Traffic density is based on the volume of traffic on major roadways within each tract (30,31).

Regionally distributed air pollution: Additionally, Sutter Health patients were assigned estimates of locally distributed air pollutant $PM_{2.5}$ according to baseline geocoded addresses. Monthly ambient data on $PM_{2.5}$ collected from air monitoring stations from 2000 to 2013 were spatially interpolated between stations using an empirical Bayesian kriging model implemented in ArcGIS to account for the uncertainty of semivariogram estimation (32). For everyone in the Sutter cohort, monthly exposure estimates relevant to the latitude/longitude coordinates of their geocoded baseline address were averaged across one year (12 months) according to the baseline date to derive the one-year average estimate of exposure at baseline. Because $PM_{2.5}$ does not capture the extent of air pollution in the Hawaiian Islands due to volcanic smog and fog ("vog"), we limited our PM data collection to the California sample.

Descriptive Statistics and Generalizability—We used descriptive statistics (frequencies and percentages) to describe the cohort and lung cancer case population overall, and by age, sex, detailed race/ethnicity, and nSES. For assessing the generalizability of the cohort to the source population, we calculated and plotted standardized mean differences using comparison estimates of background prevalence extracted from the California Health Interview Survey (CHIS) (2003–2009 and 2011–2013) (33), the Behavioral Risk Factor Surveillance System in California and Hawai'i (2000–2013) (34), and the decennial Census (2010) (35). Generalizability of lung cancer cases were assessed with covariate balance plots comparing to statewide and national (SEER-18) cancer registry data (36,37).

RESULTS

Cohort description

Our schema for the categorization of race/ethnicity (Figure 2) resulted in the disaggregation of the AANHPI group into ten single- and multi-racial/ethnic AANHPI ethnic groups. The full pooled cohort comprised 2,211,476 individuals (1,275,838 females and 935,638 males) including 49,983 Native Hawaiians, 31,506 Pacific Islanders, and 352,076 Asian Americans, representing 2.26%, 1.42%, and 15.92% of the overall cohort, respectively. There were 7,274 incident lung cancer diagnoses (3,867 females and 3,407 males) in the study period (2000–2013). These included 1,228 lung cancers among AANHPI, including 328 Native Hawaiians, 81 Pacific Islanders, and 819 Asian Americans (Tables 1 and 2).

The cohort included 889,870 females who have never smoked (69.7% of females) (Table 1). Among AANHPI, 198,208 females (79.3%) never smoked, with Asian Indian females having the highest proportion (94.7%) and Native Hawaiian females the lowest proportion (54.8%) across all racial/ethnic groups. Among females with lung cancer, the prevalence of never-smoking ranged from a low of 11.2% among non-AANHPI multiple races/ethnicities and 14.4% among Native Hawaiians to a high of 85.7% among Asian Indian and 78.7% among Chinese females (Table 1).

Among males, 545,543 (58.3%) never smoked (Table 2). Among AANHPIs, 115,573 males (63.0%) never smoked. Among males with lung cancer, the prevalence of never-smoking ranged from a low of 6.0% among Native Hawaiian to a high of 41.1% among Chinese males (Table 2).

Supplementary Tables 3a and 3b provide cohort characteristics by joint language use/ translator request. Supplementary Table 4 provides the distribution of age and detailed race/ethnicity according to nSES among the cohort.

Representativeness

Our generalizability analysis (Figure 3, Supplemental Figure 1, and Supplementary Tables 5–7) compares the pooled cohort to the demographics of the U.S. (Figure 3), and Sutter Health and Kaiser Permanente Hawai’i cohorts to the demographics of California and Hawai’i, respectively (Supplemental Figure 1). Compared to the U.S. population, the pooled cohort overrepresents the NHPI population and underrepresents non-Hispanic White, Black, and Hispanic populations. The cohort over-represents females in California, and cohort members are slightly younger on average at baseline compared to the selected representative surveys from the states of California and Hawai’i, especially for Native Hawaiian females. In the Sutter Health portion of the cohort, AANHPI proportions are generally comparable to the underlying state population, with the Sutter female population slightly overrepresenting Asian Indian (+13%), and underrepresenting Filipinx (−10%), Korean (−9%), and Vietnamese (−13%) females. The Kaiser Permanente Hawai’i female cohort slightly underrepresents NHPI (by −12%) compared to the state of Hawai’i. While not markedly different, the cohort also includes a slightly higher proportion of persons who have never smoked than the underlying population for most groups, especially among males, older adults, and the Kaiser Permanente Hawai’i cohort. Case demographics reveal a marked underrepresentation of non-Hispanic White lung cancer cases in California (−69% compared to the California Cancer Registry) but also includes enrichment for AANHPI cases (+38% compared to SEER-18). An important caveat for these findings is that our deliberate prioritization of smaller AANHPI groups in classifying cohort members' race/ethnicity differs from the Census classifications, which reflect counts of individuals identifying with "any" racial or ethnic group and thus may result in apparent over- or under-representativeness due to non-comparable classification schemes.

The geographic distribution of the cohort in California and Hawai'i are shown in Figure 4. The Sutter Health cohort, especially AANHPI females, is more concentrated in the San Francisco Bay Area and Sacramento regions of California, but has baseline residences that extend across California. Even though Kaiser Permanente Hawai'i is based on Oahu, this cohort is distributed evenly across the major islands of the state. The case population in the pooled cohort is distributed towards higher SES neighborhoods, compared to the distribution of lung cancer cases in the nation (Supplementary Tables 7a and 7b).

DISCUSSION

We have described a large-scale, EHR-based cohort inclusive of males and females of any race/ethnicity, but notably sufficient in scale to study lung cancer incidence and etiology among detailed AANHPI ethnic groups (1). By utilizing prospectively collected EHR data and integrative data analysis, this multilevel dataset includes routine clinical data from two EHRs, data on lung cancer diagnoses from state cancer registries, geospatial data on built and social neighborhood environments, and air pollution data. The study population is enriched for AANHPI groups in the U.S. while being generally representative of the target populations of California and Hawai'i. We have shown that the proportion of lung cancer diagnoses among persons who never smoked varies widely across specific AANHPI groups, with more than three-quarters of Chinese, Asian Indian, and Vietnamese female lung cancer diagnoses occurring among those who have never smoked. With ongoing work, we will estimate smoking-specific incidence of lung cancer, including incidence of histologic cell-types, according to sex and detailed race/ethnicity. With multilevel data on important known and putative risk factors, we will also leverage this linked dataset to examine multilevel etiology of lung cancer among AANHPI females who have never smoked. More broadly, this cohort and dataset illustrate a methodological approach to effectively study cancer outcomes among small populations. Our approach highlights many benefits of an integrative data analysis approach including (1) data pooling to enrich small group representation, (2) population-based registry linkages for outcome ascertainment, and (3) integration of multilevel data, each described in more detail below (38,39).

Small group representation.

Previous cancer studies in the U.S. largely considered AANHPI populations in aggregate, although smoking prevalence and proportions of lung cancer among AANHPI who have never smoked has been shown to vary widely (16). Our inclusion of multiple healthcare systems with high representation of AANHPI groups and with detailed race categories, Sutter Health and Kaiser Permanente Hawai'i, provided robust representation of AANHPI single- and multi-racial/ethnic groups. This allows us to study cancer among AANHPI groups in a way that has not been possible with prior U.S. studies and, more broadly, highlights the potential for use of EHRs to study cancer outcomes among small populations.

However, an EHR-based cohort may not be generalizable to the target populations and, as a result, studies using this type of resource may be limited in external validity (23,40,41). We aimed to address this concern with our analysis of the representativeness of the cohort, which indicated that the cohort was fairly generalizable. Representativeness, especially important when studying disease occurrence and risk factors across heterogeneous populations, is often logistically infeasible in longitudinal cohort studies (38,42). Generalizability analyses such as ours can help to establish “target validity” for EHR-based research and improve confidence in study results and subsequent interventions or policy recommendations (43,44).

Registry linkages – Outcome ascertainment.

Relying solely on EHR data for identification of incident cancer cases is problematic due to the low sensitivity of EHR-based algorithms for identifying cancer patients and the frequent migration of patients in- and out- of healthcare systems (23,45,46). Therefore we leveraged a previous linkage of Sutter Health patients to the California Cancer Registry (46) and linked the Kaiser Permanente Hawai’i patient population to the Hawai’i Tumor Registry, which allowed for higher quality, adjudicated cancer outcome ascertainment. The state cancer registries have near complete capture of cancer diagnoses in their respective states, so only individuals who moved out of the state would be lost to follow-up. To assure even greater follow-up for outcome, we required at least one record of an in-state address (California for Sutter Health and Hawai’i for Kaiser Permanente Hawai’i) during the study period. Through these registry linkages, the cohort dataset can be expanded with additional registry data to study other cancer outcomes (e.g., treatment or survival).

Integration of multilevel exposure data.

In recognition of the multilevel factors that shape cancer risk, this cohort has been geocoded to enable inclusion of data on geospatial risk factors. Current data includes baseline residential neighborhood defined as census tracts, which can be used to study neighborhood built and social environments (47–52). Linkage of the Sutter Health dataset to regional monthly air pollution estimates facilitates studies of air pollution exposure and cancer outcomes. Our use of a single address at baseline does limit the understanding of cumulative effect of geospatial exposures over time since we do not know how long individuals’ have resided at this address. Geocoding of longitudinal addresses over follow-up is possible with EHR datasets (assuming the individual remains associated with the healthcare system), though not feasible within the scope of the present study. Our use of data linkages will allow us to investigate the relative importance of many clinical and geospatial lung cancer risk factors, but these data can be limited for addressing the contribution of exposures not captured as part of routine healthcare delivery like nutrition, occupational exposures, individual socioeconomic status, and cultural factors. However, contemporary efforts to support routine capture of information on social determinants of health in the EHR will lead to increased availability of such exposures for future studies (53,54).

Use of EHR data for epidemiologic studies has considerable application and promise. This approach to cohort development is time- and cost-efficient in comparison to traditional prospective cancer cohort creation, which involves recruitment, survey development, questionnaire administration, and follow-up and retention (21,22,55). Moreover, EHR-based cohorts are particularly advantageous for assessing small populations and groups that are traditionally left out of epidemiologic studies, particularly as they are not subject to other common selection biases (including due to language barriers), non-response, attrition, and survival biases. Of note, too, the efficient approach we outline here can provide rationale for more costly prospective cohort studies able to generate consistent exposure data not available through EHR. However, as others have discussed, there are important considerations as EHR data are not collected for research purposes (21,22,55). EHR data are more limited in scope and substantial effort is required to compile and carefully curate the data while attending to the inherent biases in the use of routine healthcare data for observational research (40,56–59). We discuss below these considerations related to (1) operationalizing routinely collected EHR data for research and (2) patient privacy and institutional risk.

Data linkage and pooling– Operationalizing routinely collected EHR data for research.

Careful planning and harmonized strategies for defining and classifying clinical data elements assured consistent data extraction across Sutter Health and Kaiser Permanente EHRs and allowed us to take advantage of the full extent of the available data. This was especially important here as a unique aspect of this project was the differing nature of the partner healthcare systems. While Kaiser Permanente is a managed care provider and has an enrolled population with very little “out of plan” use not captured in the EHR, Sutter Health is a mixed payer environment with many of its patients being able to see other providers for their care. While presenting challenges in the harmonization of pooled clinical data elements, it afforded the opportunity to directly examine heterogeneity of data and data collection methods across the healthcare systems. We developed common variable definitions and, in many cases, several alternative definitions, in consideration of the availability and structure of EHR data at the two sites. Alternative definitions allowed us to specify upfront the sensitivity analyses to assess the robustness of results and assure necessary EHR data were collected without excessive repeat extractions. Coding dictionaries specifying data elements, source, timing of extraction, and category values were provided to collaborators at both healthcare systems. Moreover, because the availability of EHR data will differ across individuals with respect to healthcare utilization, we included several metrics of healthcare utilization in our data collection scheme with which to assess these potential biases in future studies of lung cancer etiology. With EHR-based research becoming more commonplace, this work may thus serve as an example for researchers pooling across system types.

Patient Privacy and Institutional Risk:

Our approach to pooling data from multiple healthcare systems carries distinct advantages for creating a sufficiently large dataset for our research purposes. It also required extensive attention to protect against data breach and adherence to use of minimum necessary protected health information (PHI) to accomplish our goals. We used data stewards for data extraction from the EHR and, except for the geocoding team, our researchers only had access to HIPAA-defined “limited” PHI. When direct identifiers were necessary (e.g., for geocoding) multiple strategies were employed to ensure safe handling of data and reduce institutional risk. For example, clinical data (from the EHR) was processed separately from patient geocodes; patient addresses and associated latitude/longitude coordinates were destroyed before combining geocode-derived exposure data (e.g., air pollution measures) with any clinical data. The development of these data safety protocols along with obtaining the appropriate approvals needed for data sharing is complicated and time-consuming. Researchers should be aware of the time and effort required and build this into their project timelines. However, once the dataset is developed, the data can be leveraged to study other cancer outcomes, especially those available in cancer registry data, for any of the racial/ethnic populations captured in the EHR.

CONCLUSION

This diverse, multilevel dataset will allow for much-needed research on lung cancer risk, especially among AANHPI female never-smokers, and serve as a critical evidence base to inform screening, research, and public health priorities in this growing population. More broadly, the innovative approach and methods used to develop this multilevel, integrated dataset can serve as an example of an effective way to conduct valuable cancer research assessing multilevel factors on cancer outcomes among small populations (60–62).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We acknowledge Dr. Aleymayehu Midekisa, Ms. Katherine Lin, and Ms. Nam Phuong Le for analytic support. We thank Dr. Brenda Hernandez with the Hawai’i Tumor Registry for providing the distribution of neighborhood socioeconomic status among Hawai’i lung cancer cases.

Financial support: National Cancer Institute (R01 CA204070; multiple PIs Gomez and Cheng). National Center for Advancing Translational Sciences (KL2TR001444; Thompson)

REFERENCES

1. Gomez SL, Chang ET, Shema SJ, Fish K, Sison JD, Reynolds P, et al. Survival following non-small cell lung cancer among Asian/Pacific Islander, Latina, and Non-Hispanic white women who have never smoked. Cancer Epidemiol Biomarkers Prev. 2011;20:545–54. [PubMed: 21239685]
2. Siegel DA, Fedewa SA, Henley SJ, Pollack LA, Jemal A. Proportion of Never Smokers Among Men and Women With Lung Cancer in 7 US States. JAMA Oncol. 2020;
3. Subramanian J, Govindan R. Lung cancer in never smokers: a review. J Clin Oncol. 2007;25:561–70. [PubMed: 17290066]
4. Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers--a different disease. Nat Rev Cancer. 2007;7:778–90. [PubMed: 17882278]
5. Sisti J, Boffetta P. What proportion of lung cancer in never-smokers can be attributed to known risk factors? Int J Cancer. 2012;131:265–75. [PubMed: 22322343]
6. The Rise of Asian Americans. Washington, DC: Pew Researach Center; 2012 6

7. Ponce NA, Tseng W, Ong P, Shek YL, Ortiz S, Gatchell M. The State of Asian American, Native Hawaiian and Pacific Islander Health in California Report. 2009 [cited 2020 Jul 24]; Available from: https://escholarship.org/uc/item/3s89c1cm

8. A Community of Contrasts: Native Hawaiians and Pacific Islanders in the United States [Internet]. Los Angeles, CA: Asian Americans Advancing Justice _LA; 2014. Available from: https://www.advancingjustice-la.org/what-we-do/policy-and-research/demographic-research/community-contrasts-native-hawaiians-and-pacific

9. Moy KL, Sallis JF, Trinidad DR, Ice CL, McEligot AJ. Health Behaviors of Native Hawaiian and Pacific Islander Adults in California. Asia Pac J Public Health. 2012;24:961–9. [PubMed: 22426559]

10. The Asian Population: 2010 [Internet]. Washington, DC: United States Department of Commrece, Economics, and Statistics Administration. United States Census Bureau; 2012 Mar page 24. Available from: https://www.census.gov/prod/cen2010/briefs/c2010br-11.pdf

11. The Native Hawaiian and Other Pacific Islander Population: 2010 [Internet]. Washington, DC: United States Department of Commrece, Economics, and Statistics Administration. United States Census Bureau; 2012 5 Available from: https://www.census.gov/prod/cen2010/briefs/c2010br-12.pdf

12. Asian/Pacific American Heritage Month: May 2014 [Internet]. Washington, DC: United States Department of Commrece, Economics, and Statistics Administration. United States Census Bureau; 2014 4. Available from: https://www.census.gov/content/dam/Census/newsroom/facts-for-features/2014/cb14-ff13_asian.pdf

13. Srinivasan S, Guillermo T. Toward improved health: disaggregating Asian American and Native Hawaiian/Pacific Islander data. Am J Public Health. 2000;90:1731–4. [PubMed: 11076241]

14. Nguyen AB, Chawla N, Noone A-M, Srinivasan S. Disaggregated data and beyond: future queries in cancer control research. Cancer Epidemiol Biomarkers Prev. 2014;23:2266–72. [PubMed: 25368401]

15. Gomez SL, Glaser SL, Horn-Ross PL, Cheng I, Quach T, Clarke CA, et al. Cancer Research in Asian American, Native Hawaiian, and Pacific Islander Populations: Accelerating Cancer Knowledge by Acknowledging and Leveraging Heterogeneity. Cancer Epidemiol Biomarkers Prev. 2014;23:2202–5. [PubMed: 25368394]

16. Srinivasan S, Moser RP, Willis G, Riley W, Alexander M, Berrigan D, et al. Small Is Essential: Importance of Subpopulation Research in Cancer Control. Am J Public Health. 2015;105:S371–3. [PubMed: 25905825]

17. National Academies of Sciences, Engineering, and Medicine, Health and Medicine Division, Board on Health Care Services, Division of Behavioral and Social Sciences and Education, Committee on National Statistics. Improving Health Research on Small Populations: Proceedings of a Workshop [Internet]. Washington (DC): National Academies Press (US); 2018 [cited 2021 Feb 16]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK499549/

18. Atienza AA, Serrano KJ, Riley WT, Moser RP, Klein WM. Advancing Cancer Prevention and Behavior Theory in the Era of Big Data. J Cancer Prev. 2016;21:201–6. [PubMed: 27722147]

19. Allen J, Inder KJ, Lewin TJ, Attia JR, Kay-Lambkin FJ, Baker AL, et al. Integrating and extending cohort studies: lessons from the eXtending Treatments, Education and Networks in Depression (xTEND) study. BMC Med Res Methodol. 2013;13:122. [PubMed: 24093910]

20. Curran PJ, Hussong AM. Integrative data analysis: the simultaneous analysis of multiple data sets. Psychol Methods. 2009;14:81–100. [PubMed: 19485623]

21. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-World Evidence - What Is It and What Can It Tell Us? N Engl J Med. 2016;375:2293–7. [PubMed: 27959688]

22. Khozin S, Blumenthal GM, Pazdur R. Real-world Data for Clinical Evidence Generation in Oncology. J Natl Cancer Inst. 2017;109.

23. Thompson C, Jin A, Luft HS, Lichtensztajn DY, Allen L, Liang S-Y, et al. Population-based registry linkages to improve validity of electronic health record-based cancer research. Cancer Epidemiol Biomarkers Prev [Internet]. American Association for Cancer

Research; 2020 [cited 2020 Jul 24]; Available from: https://cebp.aacrjournals.org/content/early/2020/02/15/1055-9965.EPI-19-0882

24. Lewis DR, Check DP, Caporaso NE, Travis WD, Devesa SS. US lung cancer trends by histologic type. Cancer. 2014;120:2883–92. [PubMed: 25113306]

25. Census Questionnaires (2000 and 1990) [Internet]. [cited 2020 Jul 24]. Available from: https://www.census.gov/dmd/www/2000quest.html

26. Suinn RM, Rickard-Figueroa K, Lew S, Vigil P. The Suinn-Lew Asian Self-Identity Acculturation Scale: An Initial Report. Educational and Psychological Measurement. SAGE Publications Inc; 1987;47:401–7.

27. ArcGIS | Mapping & Analytics Platform [Internet]. Rdlands, CA: Esri; [cited 2020 Jul 30]. Available from: https://www.esri.com/en-us/arcgis/about-arcgis/overview

28. Yost K, Perkins C, Cohen R, Morris C, Wright W. Socioeconomic status and breast cancer incidence in California for different race/ethnic groups. Cancer Causes Control. 2001;12:703–11. [PubMed: 11562110]

29. Yang J, Schupp CW, Harrati A, Clarke C, Keegan T, Gomez S. Developing an area-based socioeconomic measure from American Community Survey data. Fremont, California: Cancer Prevention Institute of California; 2014.

30. Gomez SL, Clarke CA, Shema SJ, Chang ET, Keegan THM, Glaser SL. Disparities in Breast Cancer Survival Among Asian Women by Ethnicity and Immigrant Status: A Population-Based Study. Am J Public Health. 2010;100:861–9. [PubMed: 20299648]

31. DeRouen MC, Schupp CW, Koo J, Yang J, Hertz A, Shariff-Marco S, et al. Impact of individual and neighborhood factors on disparities in prostate cancer survival. Cancer Epidemiol. 2018. page 1–11.

32. Pilz J, Spöck G. Why do we need and how should we implement Bayesian kriging methods. Stoch Environ Res Risk Assess. 2008;22:621–32.

33. California Health Interview Survey | UCLA Center for Health Policy Research [Internet]. UCLA Center for Health Policy Research, UCLA Fielding School of Public Health. [cited 2020 Jul 30]. Available from: http://healthpolicy.ucla.edu/chis/Pages/default.aspx

34. Behavioral Risk Factor Surveillance System - Survey Data & Documentation [Internet]. Centers for Disease Control and Prevention. 2019 [cited 2020 Jul 30]. Available from: https://www.cdc.gov/brfss/data_documentation/index.htm

35. Bureau UC. Decennial Census Datasets [Internet]. The United States Census Bureau. [cited 2020 Aug 6]. Available from: https://www.census.gov/programs-surveys/decennial-census/data/datasets.html

36. SEER Incidence Database - SEER Data & Software [Internet]. SEER. [cited 2020 Aug 6]. Available from: https://seer.cancer.gov/data/index.html

37. SEER*Stat Software [Internet]. SEER. [cited 2020 Aug 6]. Available from: https://seer.cancer.gov/seerstat/index.html

38. Rothman KJ, Gallacher JEJ, Hatch EE. Why representativeness should be avoided. International Journal of Epidemiology. 2013;42:1012–4. [PubMed: 24062287]

39. Nohr EA, Olsen J. Commentary: Epidemiologists have debated representativeness for more than 40 years--has the time come to move on? International Journal of Epidemiology. 2013;42:1016–7. [PubMed: 24062289]

40. Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research - PubMed. [cited 2020 Jul 24]; Available from: https://pubmed.ncbi.nlm.nih.gov/24916006/

41. Gomez SL, Shariff-Marco S, Von Behren J, Kwan ML, Kroenke CH, Keegan THM, et al. Representativeness of breast cancer cases in an integrated health care delivery system. BMC Cancer. 2015;15:688. [PubMed: 26467773]

42. Stang A, Jöckel K-H. Avoidance of representativeness in presence of effect modification. Int J Epidemiol. 2014;43:630–1. [PubMed: 24408970]

43. Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing study results: a potential outcomes perspective. Epidemiology. 2017;28:553–61. [PubMed: 28346267]

44. Westreich D, Edwards JK, Lesko CR, Cole SR, Stuart EA. Target Validity and the Hierarchy of Study Designs. American Journal of Epidemiology. 2019;188:438–43. [PubMed: 30299451]

45. Weber SC, Seto T, Olson C, Kenkare P, Kurian AW, Das AK. Oncoshare: lessons learned from building an integrated multi-institutional database for comparative effectiveness research. AMIA Annu Symp Proc. 2012;2012:970–8. [PubMed: 23304372]

46. Thompson CA, Kurian AW, Luft HS. Linking electronic health records to better understand breast cancer patient pathways within and between two health systems. EGEMS (Wash DC). 2015;3:1127. [PubMed: 25992389]

47. Anjum Hajat, Diez-Roux Ana V, Adar Sara D, Auchincloss Amy H, Lovasi Gina S, O'Neill Marie S, et al. Air Pollution and Individual and Neighborhood Socioeconomic Status: Evidence from the Multi-Ethnic Study of Atherosclerosis (MESA). Environmental Health Perspectives. Environmental Health Perspectives; 2013;121:1325–33. [PubMed: 24076625]

48. Dubowitz T, Ghosh-Dastidar M, Eibner C, Slaughter ME, Fernandes M, Whitsel EA, et al. The Women's Health Initiative: The Food Environment, Neighborhood Socioeconomic Status, BMI, and Blood Pressure. Obesity. 2012;20:862–71. [PubMed: 21660076]

49. Kish JK, Yu M, Percy-Laurry A, Altekruse SF. Racial and Ethnic Disparities in Cancer Survival by Neighborhood Socioeconomic Status in Surveillance, Epidemiology, and End Results (SEER) Registries. JNCI Monographs. 2014;2014:236–43.

50. Shariff-Marco S, DeRouen MC, Yang J, Jain J, Nelson DO, Weden MM, et al. Neighborhood archetypes and breast cancer survival in California. Ann Epidemiol. 2021;

51. DeRouen MC, Hu L, McKinley M, Gali K, Patel M, Clarke C, et al. Incidence of lung cancer histologic cell-types according to neighborhood factors: A population based study in California. PLoS One. 2018. page e0197146.

52. Patel MI, McKinley M, Cheng I, Haile R, Wakelee H, Gomez SL. Lung cancer incidence trends in California by race/ethnicity, histology, sex, and neighborhood socioeconomic status: An analysis spanning 28 years. Lung Cancer. 2017;108:140–9. [PubMed: 28625626]

53. Cantor MN, Thorpe L. Integrating Data On Social Determinants Of Health Into Electronic Health Records. Health Affairs. Health Affairs; 2018;37:585–90. [PubMed: 29608369]

54. Hatef E, Predmore Z, Lasser EC, Kharrazi H, Nelson K, Curtis I, et al. Integrating social and behavioral determinants of health into patient care and population health at Veterans Health Administration: a conceptual framework and an assessment of available individual and population level data sources and evidence-based measurements. AIMS Public Health. 2019;6:209–24. [PubMed: 31637271]

55. Mahajan R. Real world data: Additional source for making clinical decisions. Int J Appl Basic Med Res. 2015;5:82. [PubMed: 26097811]

56. Bower JK, Patel S, Rudy JE, Felix AS. Addressing Bias in Electronic Health Record-Based Surveillance of Cardiovascular Disease Risk: Finding the Signal Through the Noise. Curr Epidemiol Rep. 2017;4:346–52. [PubMed: 31223556]

57. Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. J Med Internet Res. 2018;20:e185. [PubMed: 29844010]

58. Goldstein BA, Bhavsar NA, Phelan M, Pencina MJ. Controlling for Informed Presence Bias Due to the Number of Health Encounters in an Electronic Health Record. Am J Epidemiol. 2016;184:847–55. [PubMed: 27852603]

59. Weber GM, Adams WG, Bernstam EV, Bickel JP, Fox KP, Marsolo K, et al. Biases introduced by filtering electronic health records for patients with "complete data." J Am Med Inform Assoc. 2017;24:1134–41. [PubMed: 29016972]

60. Lynch SM, Rebbeck TR. Bridging the gap between biologic, individual, and macroenvironmental factors in cancer: a multilevel approach. Cancer Epidemiol Biomarkers Prev. 2013;22:485–95. [PubMed: 23462925]

61. Warnecke RB, Oh A, Breen N, Gehlert S, Paskett E, Tucker KL, et al. Approaching Health Disparities From a Population Perspective: The National Institutes of Health Centers for Population Health and Health Disparities. Am J Public Health. 2008;98:1608–15. [PubMed: 18633099]

62. Alvidrez J, Castille D, Laude-Sharp M, Rosario A, Tabor D. The National Institute on Minority Health and Health Disparities Research Framework. Am J Public Health. 2019;109:S16–20. [PubMed: 30699025]

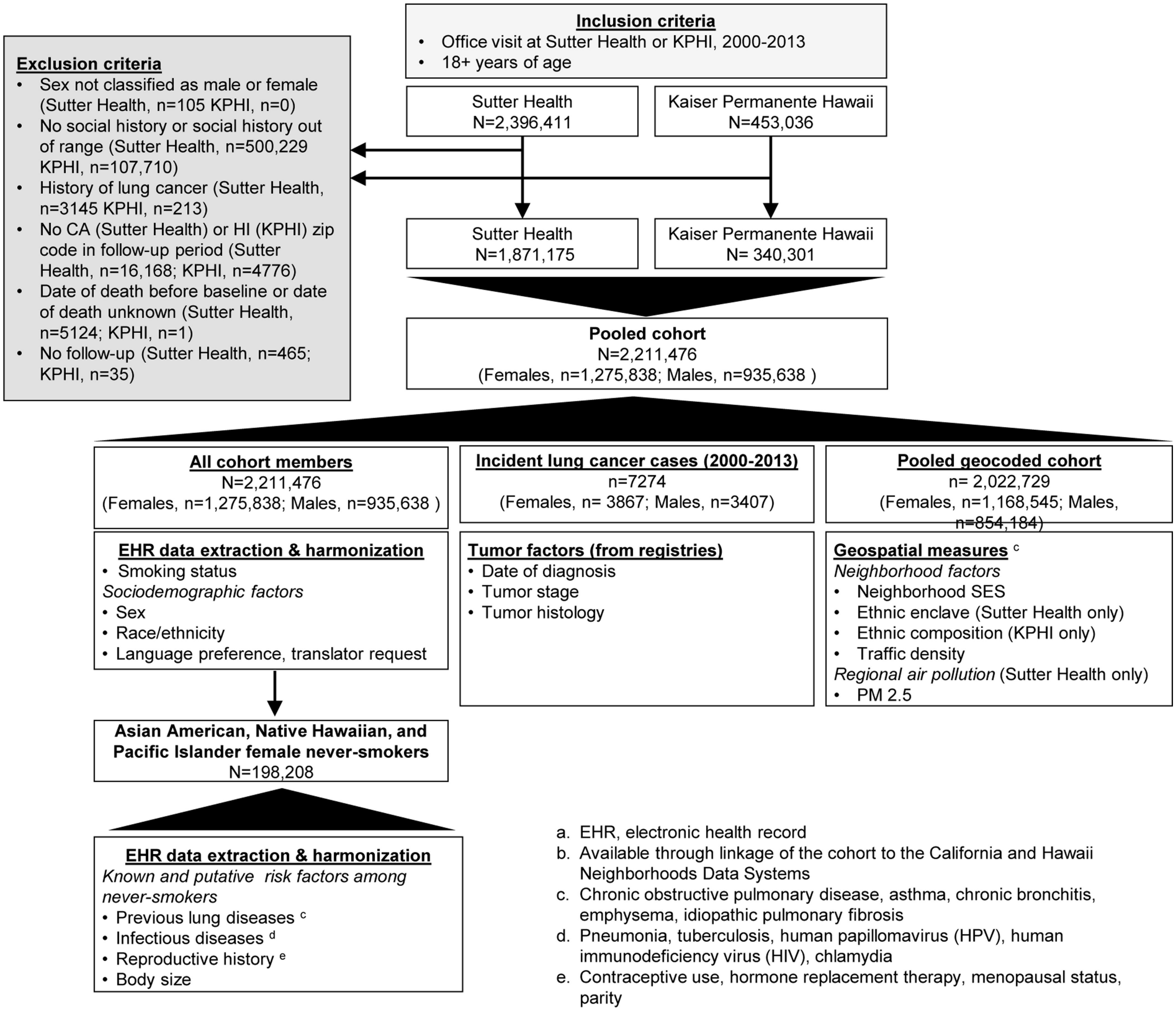
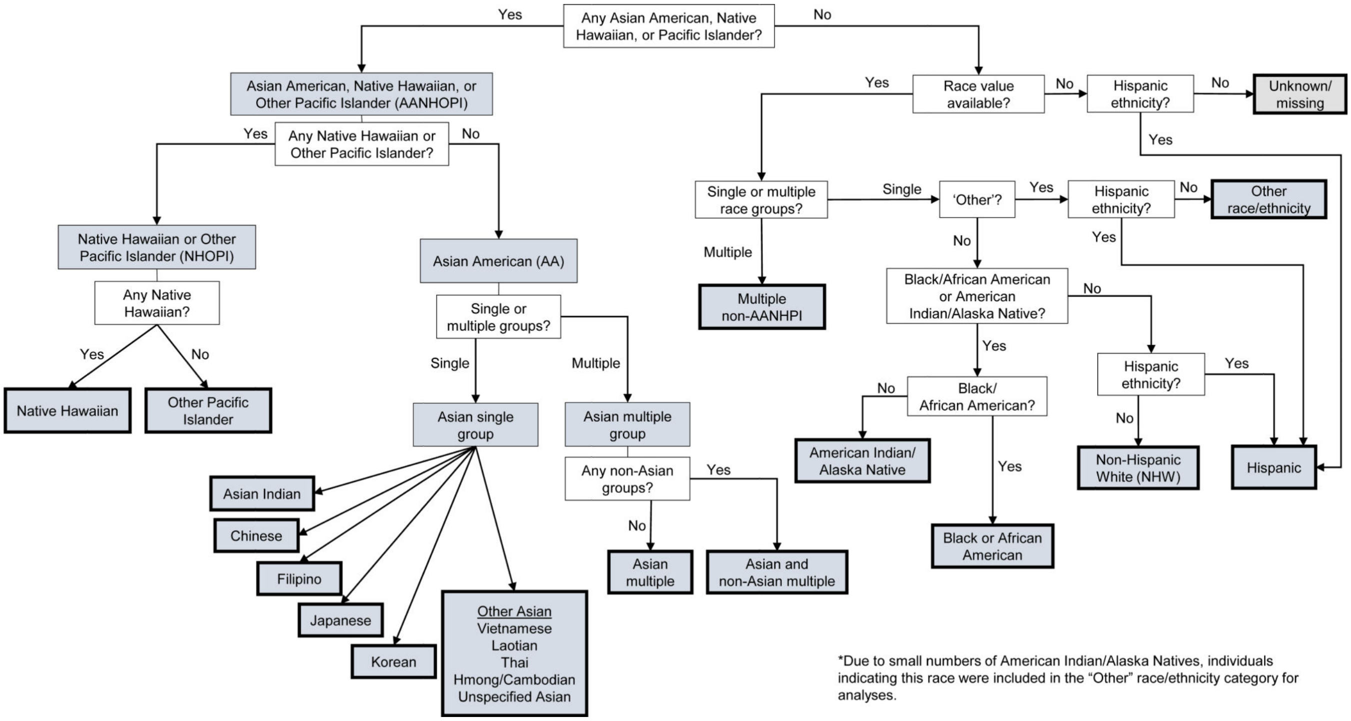

Inclusion criteria
• Office visit at Sutter Health or KPHI, 2000-2013
• 18+ years of age
Sutter Health N=2,396,411
Kaiser Permanente Hawaii N=453,036
Exclusion criteria
• Sex not classified as male or female (Sutter Health, n=105 KPHI, n=0)
• No social history or social history out of range (Sutter Health, n=500,229 KPHI, n=107,710)
• History of lung cancer (Sutter Health, n=3145 KPHI, n=213)
• No CA (Sutter Health) or HI (KPHI) zip code in follow-up period (Sutter Health, n=16,168; KPHI, n=4776)
• Date of death before baseline or date of death unknown (Sutter Health, n=5124; KPHI, n=1)
• No follow-up (Sutter Health, n=465; KPHI, n=35)
Sutter Health N=1,871,175
Kaiser Permanente Hawaii N= 340,301
Pooled cohort N=2,211,476 (Females, n=1,275,838; Males, n=935,638)
All cohort members N=2,211,476 (Females, n=1,275,838; Males, n=935,638)
Incident lung cancer cases (2000-2013) n=7274 (Females, n= 3867; Males, n=3407)
Pooled geocoded cohort n= 2,022,729 (Females, n=1,168,545; Males, n=854,184)
EHR data extraction & harmonization
• Smoking status
Sociodemographic factors
• Sex
• Race/ethnicity
• Language preference, translator request
Tumor factors (from registries)
• Date of diagnosis
• Tumor stage
• Tumor histology
Geospatial measures c
Neighborhood factors
• Neighborhood SES
• Ethnic enclave (Sutter Health only)
• Ethnic composition (KPHI only)
• Traffic density
Regional air pollution (Sutter Health only)
• PM 2.5
Asian American, Native Hawaiian, and Pacific Islander female never-smokers N=198,208
EHR data extraction & harmonization
Known and putative risk factors among never-smokers
• Previous lung diseases c
• Infectious diseases d
• Reproductive history e
• Body size
a. EHR, electronic health record
b. Available through linkage of the cohort to the California and Hawaii Neighborhoods Data Systems
c. Chronic obstructive pulmonary disease, asthma, chronic bronchitis, emphysema, idiopathic pulmonary fibrosis
d. Pneumonia, tuberculosis, human papillomavirus (HPV), human immunodeficiency virus (HIV), chlamydia
e. Contraceptive use, hormone replacement therapy, menopausal status, parity

Figure 1. Cohort specification and multilevel data integration, Sutter Health and Kaiser Permanente Hawai'i Lung Cancer Cohort 2000–2013.

Individuals eligible for cohort inclusion were female or male adults (≥18 years old) who were seen in-person by a provider at Sutter Health or Kaiser Permanente Hawai'i between January 1, 2000 and December 31, 2013 (the study end date) for an initial sample of n=2,396,411 from Sutter Health and n=453,036 from Kaiser Permanente Hawai'i. Individuals' first in-person encounter was classified as their baseline visit. We excluded individuals for having 1) missing sex (Sutter n=105 and Kaiser Permanente N=0); 2) no data collected for social history (Sutter n=412,214; Kaiser Permanente n=94,449) or social history collected outside of the study period (Sutter n=88,015; Kaiser Permanente n=13,261); 3) pre-baseline history of lung cancer as determined by searching EHR records (ICD-9 codes; 162.0, 162.2, 162.3, 162.4, 162.5, 162.8, 162.9) or California (Sutter Health) or Hawai'i (Kaiser Permanente) cancer registry linkage (SEER site recode ICD-O-3/WHO 2008 = 22030) (Sutter Health, n=3,145; Kaiser Permanente, n=213); 4) no evidence of a California (Sutter Health) or Hawai'i (Kaiser Permanente) address, as determined by billing zip code (Sutter Health, n=16,168; Kaiser Permanente, n=4,776); 5) date of death before the baseline date (Sutter Health n=56; Kaiser Permanente n=1), 6) date of death unknown for deceased non-cases only (Sutter Health n=5,068; Kaiser Permanente n=0), and 7) no follow-up (Sutter Health n=465; Kaiser Permanente n=35). The final pooled cohort includes N=2,211,476 (by sex: n=935,638 males, n=1,275,838 females; by system: Sutter n=1,871,175, Kaiser Permanente n=340,301).



*Due to small numbers of American Indian/Alaska Natives, individuals indicating this race were included in the "Other" race/ethnicity category for analyses.

Figure 2. Approach to categorization of detailed race/ethnicity and multi-racial/ethnic groups, Sutter Health and Kaiser Permanente Hawai'i Lung Cancer Cohort 2000–2013.

Boxes outlined in bold indicate final race/ethnicity categories.

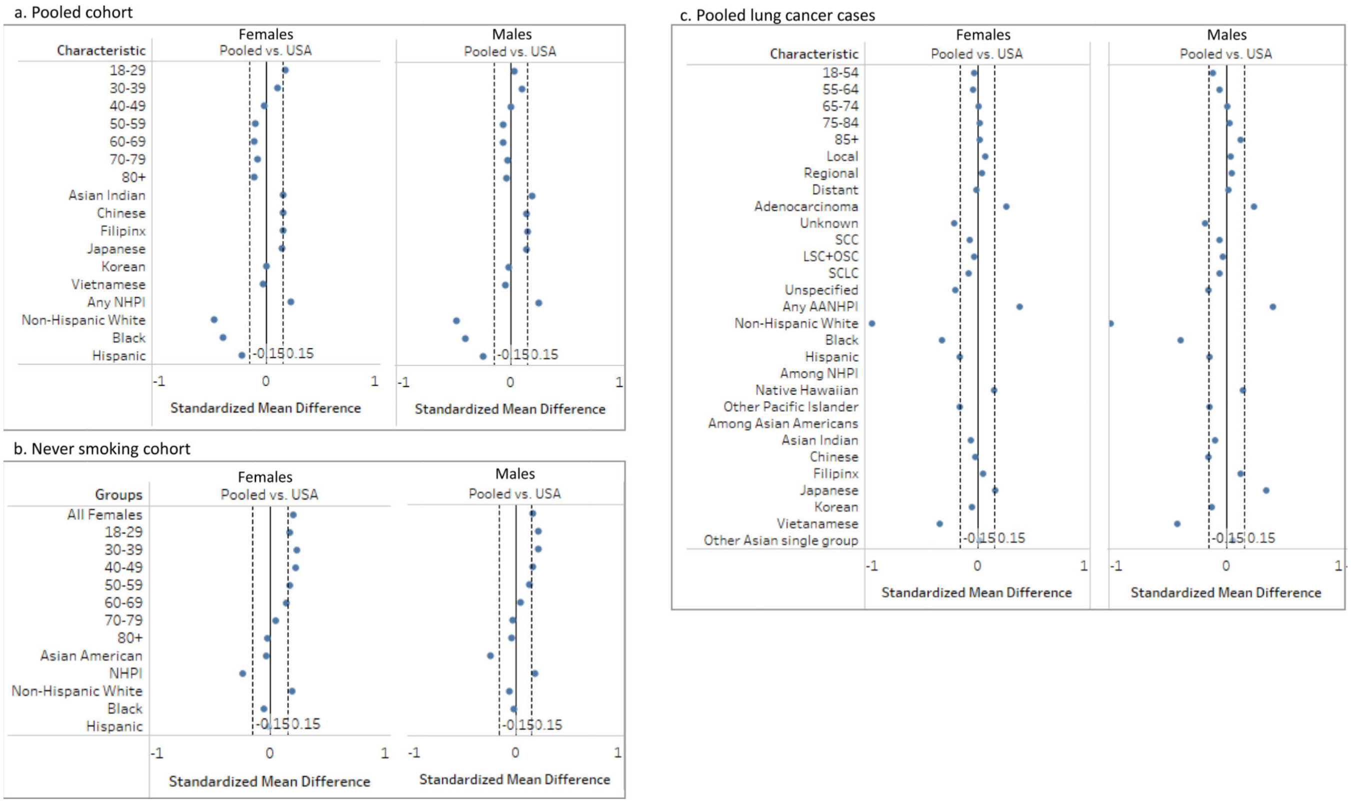


Figure 3. Representativeness of the Sutter Health and Kaiser Permanente Hawai'i Lung Cancer Cohort, 2000–2013.

(a) Distribution of females and males in the pooled cohort by age and race/ethnicity are compared to the 2010 US Census. (b) Proportion of never smoking females and males by age and race/ethnicityare compared to U.S. Behavioral Risk Factors Surveillance System (BRFSS) data.. (c) Demographic (age, race/ethnicity) and tumor (stage, histology) characteristics for female and male lung cancer cases in the pooled cohort were compared to SEER-18. Standardized mean differences (SMD), which may not be representative of the reference population when they are beyond +/− 15%, were calculated as $SMD = \frac{P_s - P_r}{\sqrt{\frac{\left(P_S^*(1 - P_S)\right) + \left(P_Y^*(1 - P_r)\right)}{2}}}$ where P_S is the study proportion and P_r is the reference population proportion for each variable category.

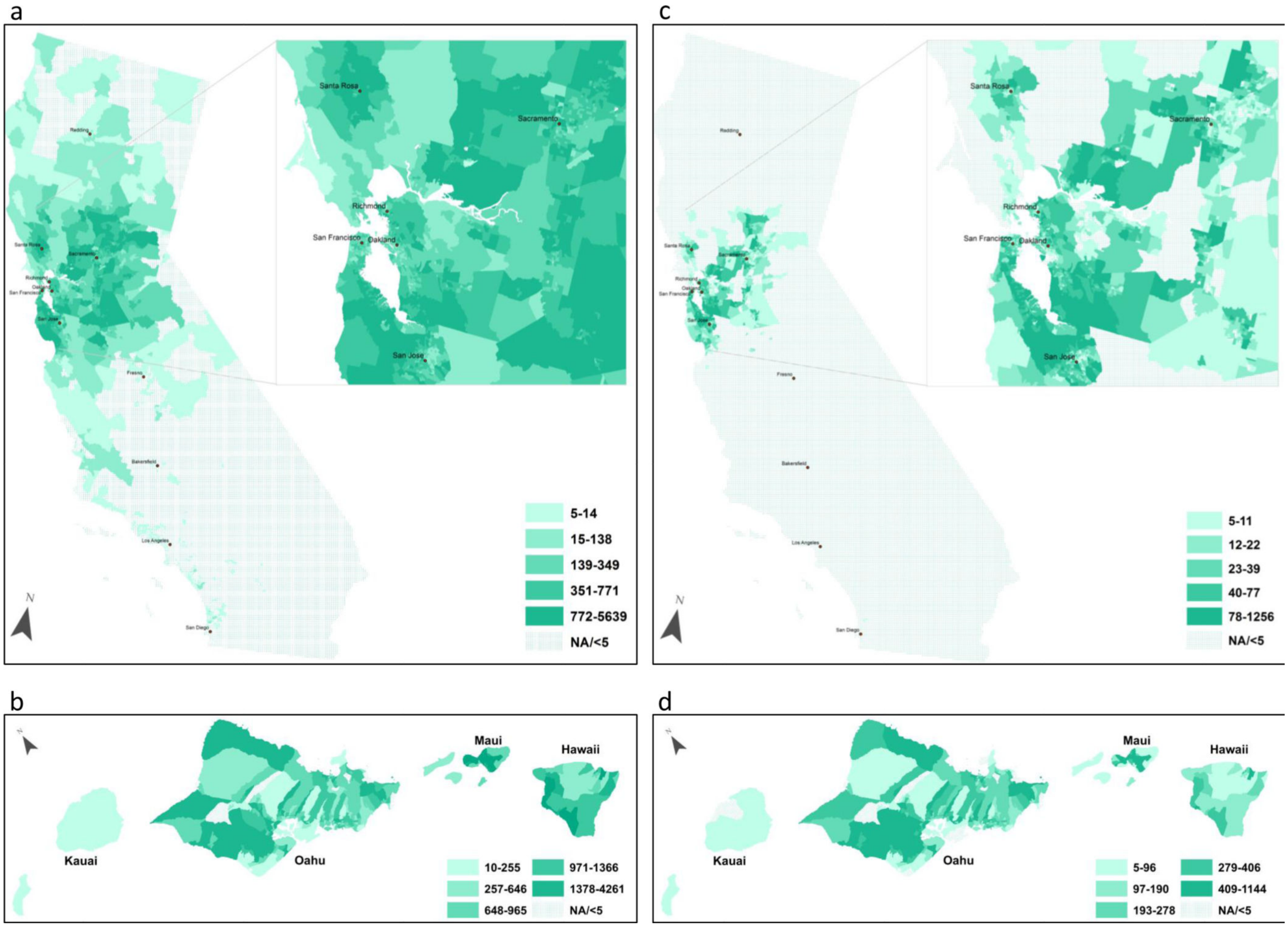


Figure 4. Geographic distribution of addresses by census tract of residence at baseline among individuals in the Sutter Health and Kaiser Permanente Hawai'i Lung Cancer Cohort 2000–2013

in (a) California and (b) Hawai'i and distribution of addresses of residence at baseline among Asian American, Native Hawaiian, and Pacific Islander females in the LCINF Lung Cancer Cohort 2000–2013 in (c) California and (d) Hawai'i.

Table 1.

Pooled cohort and lung cancer case characteristics overall and by smoking status among females, 2000–2013

	Female Cohort								**Female Incident Lung Cancer Cases**							
			Smoking Status								**Smoking Status**					
Characteristic	**Total (N=1,275,838)**		**Never (n=889,870)**		**Ever (n=332,824)**		**Unknown (n=53,144)**		**Total (N=3,867)**		**Never (n=884)**		**Ever (n=2,413)**		**Unknown (n=570)**	
	N	**(%)[d]**	**N**	**(%)**	**N**	**(%)**	**N**	**(%)**	**N**	**(%)[d]**	**N**	**(%)**	**N**	**(%)**	**N**	**(%)**
Age at baseline/diagnosis[a]																
18–29	360,091	(28.2)	272,851	(75.8)	72,748	(20.2)	14,492	(4.0)	6	(0.2)	~	~	~	~	~	~
30–39	264,467	(20.7)	198,435	(75.0)	55,231	(20.9)	10,801	(4.1)	40	(1.0)	23	(57.5)	14	(35.0)	~	~
40–49	223,175	(17.5)	154,146	(69.1)	60,835	(27.3)	8,194	(3.7)	146	(3.8)	55	(37.7)	64	(43.8)	27	(18.5)
50–59	184,312	(14.4)	116,551	(63.2)	60,515	(32.8)	7,246	(3.9)	483	(12.5)	107	(22.2)	300	(62.1)	76	(15.7)
60–69	121,471	(9.5)	71,997	(59.3)	44,364	(36.5)	5,110	(4.2)	1073	(27.7)	217	(20.2)	686	(63.9)	170	(15.8)
70–79	73,848	(5.8)	44,193	(59.8)	25,839	(35.0)	3,816	(5.2)	1237	(32.0)	232	(18.8)	819	(66.2)	186	(15.0)
80+	48,474	(3.8)	31,697	(65.4)	13,292	(27.4)	3,485	(7.2)	882	(22.8)	246	(27.9)	529	(60.0)	107	(12.1)
Race/Ethnicity																
Any AANHPI	250,053	(19.6)	198,208	(79.3)	45,939	(18.4)	5,906	(2.4)	613	(15.9)	235	(38.3)	300	(48.9)	78	(12.7)
Any NHPI	43,409	(3.4)	25,139	(57.9)	17,488	(40.3)	782	(1.8)	201	(5.2)	31	(15.4)	147	(73.1)	23	(11.4)
Native Hawaiian[b]	26,760	(2.1)	14,658	(54.8)	11,809	(44.1)	293	(1.1)	160	(4.1)	23	(14.4)	121	(75.6)	16	(10.0)
Pacific Islander[c]	16,649	(1.3)	10,481	(63.0)	5,679	(34.1)	489	(2.9)	41	(1.1)	8	(19.5)	26	(63.4)	7	(17.1)
Asian (single or multiple)	206,644	(16.2)	173,069	(83.8)	28,451	(13.8)	5,124	(2.5)	412	(10.7)	204	(49.5)	153	(37.1)	55	(13.3)
Asian (single group)	171,531	(13.4)	146,518	(85.4)	20,617	(12.0)	4,396	(2.6)	296	(7.7)	158	(53.4)	103	(34.8)	35	(11.8)
Asian Indian	37,434	(2.9)	35,458	(94.7)	924	(2.5)	1,052	(2.8)	7	(0.2)	6	(85.7)	0	(0.0)	~	~
Chinese	38,950	(3.1)	35,622	(91.5)	2,360	(6.1)	968	(2.5)	75	(1.9)	59	(78.7)	8	(10.7)	8	(10.7)
Japanese	19,804	(1.6)	14,453	(73.0)	4,988	(25.2)	363	(1.8)	74	(1.9)	18	(24.3)	39	(52.7)	17	(23.0)
Filipinx	36,539	(2.9)	28,841	(78.9)	7,011	(19.2)	687	(1.9)	80	(2.1)	42	(52.5)	34	(42.5)	~	~
Korean	6,611	(0.5)	4,887	(73.9)	1,542	(23.3)	182	(2.8)	21	(0.5)	6	(28.6)	13	(61.9)	~	~
Vietnamese	4,421	(0.3)	4,012	(90.7)	317	(7.2)	92	(2.1)	5	(0.1)	5	(100)	0	(0.0)	0	(0.0)
Other Asian	27,772	(2.2)	23,245	(83.7)	3,475	(12.5)	1,052	(3.8)	34	(0.9)	22	(64.7)	9	(26.5)	~	~

	Female Cohort								Female Incident Lung Cancer Cases							
			Smoking Status								Smoking Status					
Characteristic	Total (N=1,275,838)		Never (n=889,870)		Ever (n=332,824)		Unknown (n=53,144)		Total (N=3,867)		Never (n=884)		Ever (n=2,413)		Unknown (n=570)	
	N	(%)[d]	N	(%)	N	(%)	N	(%)	N	(%)[d]	N	(%)	N	(%)	N	(%)
Asian (multiple group)	35,113	(2.8)	26,551	(75.6)	7,834	(22.3)	728	(2.1)	116	(3.0)	46	(39.7)	50	(43.1)	20	(17.2)
Asian only multiple	10,338	(0.8)	8,461	(81.8)	1,621	(15.7)	256	(2.5)	16	(0.4)	12	(75.0)	~	~	0	(0.0)
Asian & non-Asian	24,775	(1.9)	18,090	(73.0)	6,213	(25.1)	472	(1.9)	100	(2.6)	34	(34.0)	46	(46.0)	20	(20.0)
Non-Hispanic White	535,007	(41.9)	348,492	(65.1)	169,660	(31.7)	16,855	(3.2)	1,489	(38.5)	306	(20.6)	991	(66.6)	192	(12.9)
Black	36,640	(2.9)	23,995	(65.5)	11,493	(31.4)	1,152	(3.1)	91	(2.4)	13	(14.3)	69	(75.8)	9	(9.9)
Hispanic	103,759	(8.1)	79,235	(76.4)	20,835	(20.1)	3,689	(3.6)	81	(2.1)	31	(38.3)	41	(50.6)	9	(11.1)
Non-AANHPI multiple	39,434	(3.1)	26,055	(66.1)	12,173	(30.9)	1,206	(3.1)	125	(3.2)	14	(11.2)	80	(64.0)	31	(24.8)
Other (incl AIAN)	26,159	(2.1)	19,220	(73.5)	5,975	(22.8)	964	(3.7)	23	(0.6)	~	~	14	(60.9)	5	(21.7)
Unknown	284,786	(22.3)	194,665	(68.4)	66,749	(23.4)	23,372	(8.2)	1,445	(37.4)	281	(19.4)	918	(63.5)	246	(17.0)

[a] . Age at baseline among cohort, age at diagnosis among cases.

[b] . Individuals indicating *any* Native Hawaiian, even if also indicating other races/ethnicities, are categorized as ‘Native Hawaiian’.

[c] . Pacific Islander, not indicating Native Hawaiian

[d] . Column percentages are provided in ‘Total’ columns. Percentages totaling to 100% in ‘Total’ column are mutually exclusive categories of Any AANHPI, Non-Hispanic White, Black, Hispanic, Non-AANHPI multiple, Other (incl AIAN), and Unknown. All other columns with proportions present row percentages.

[e] . The symbol ‘~’ indicates censoring due to low numbers (<5 individuals).

[f] . AANHPI, Asian American, Native Hawaiian, and Pacific Islander; NHPI, Native Hawaiian and Pacific Islander; AIAN, American Indian and Alaska Native.

Table 2.

Pooled cohort and lung cancer case characteristics overall and by smoking status among males, 2000–2013

Characteristic	Male Cohort								Male Incident Lung Cancer Cases							
	Total (N=935,638)		Smoking Status						Total (N=3,407)		Smoking Status					
			Never (n=545,543)		Ever (n=350,708)		Unknown (n=39,387)				Never (n=429)		Ever (n=2,463)		Unknown (n=515)	
	N	(%)[d]	N	(%)	N	(%)	N	(%)	N	(%)[d]	N	(%)	N	(%)	N	(%)
Age at baseline/diagnosis [a]																
18–29	229,314	(24.5)	150,322	(65.6)	69,393	(30.3)	9,599	(4.2)	8	(0.2)	5	(62.5)	~	~	~	~
30–39	201,532	(21.5)	131,728	(65.4)	62,333	(30.9)	7,471	(3.7)	15	(0.4)	10	(66.7)	~	~	~	~
40–49	179,262	(19.2)	109,640	(61.2)	62,573	(34.9)	7,049	(3.9)	90	(2.6)	19	(21.1)	59	(65.6)	12	(13.3)
50–59	147,078	(15.7)	77,277	(52.5)	63,744	(43.3)	6,057	(4.1)	413	(12.1)	58	(14.0)	293	(70.9)	62	(15.0)
60–69	94,829	(10.1)	42,114	(44.4)	48,448	(51.1)	4,267	(4.5)	900	(26.4)	87	(9.7)	678	(75.3)	135	(15.0)
70–79	55,128	(5.9)	22,263	(40.4)	29,917	(54.3)	2,948	(5.3)	1159	(34.0)	121	(10.4)	853	(73.6)	185	(16.0)
80+	28,495	(3.0)	12,199	(42.8)	14,300	(50.2)	1,996	(7.0)	822	(24.1)	129	(15.7)	574	(69.8)	119	(14.5)
Race/Ethnicity																
Any AANHPI	183,512	(19.6)	115,573	(63.0)	64,598	(35.2)	3,341	(1.8)	615	(18.1)	97	(15.8)	414	(67.3)	104	(16.9)
Any NHPI	38,080	(4.1)	19,225	(50.5)	18,494	(48.6)	361	(0.9)	208	(6.1)	14	(6.7)	155	(74.5)	39	(18.8)
Native Hawaiian[b]	23,223	(2.5)	11,565	(49.8)	11,466	(49.4)	192	(0.8)	168	(4.9)	10	(6.0)	123	(73.2)	35	(20.8)
Pacific Islander[c]	14,857	(1.6)	7,660	(51.6)	7,028	(47.3)	169	(1.1)	40	(1.2)	~	~	32	(80.0)	~	~
Asian (single or multiple group)	145,432	(15.5)	96,348	(66.2)	46,104	(31.7)	2,980	(2.0)	407	(11.9)	83	(20.4)	259	(63.6)	65	~
Asian (single group)	124,926	(13.4)	84,033	(67.3)	38,257	(30.6)	2,636	(2.1)	307	(9.0)	65	(21.2)	195	(63.5)	47	(15.3)
Asian Indian	34,947	(3.7)	27,100	(77.5)	7,027	(20.1)	820	(2.3)	9	(0.3)	~	~	5	(55.6)	~	~
Chinese	27,361	(2.9)	20,918	(76.5)	5,918	(21.6)	525	(1.9)	56	(1.6)	23	(41.1)	27	(48.2)	6	(10.7)
Japanese	14,271	(1.5)	7,860	(55.1)	6,236	(43.7)	175	(1.2)	86	(2.5)	13	(15.1)	56	(65.1)	17	(19.8)
Filipinx	26,291	(2.8)	14,114	(53.7)	11,831	(45.0)	346	(1.3)	103	(3.0)	16	(15.5)	72	(69.9)	15	(14.6)
Korean	3,574	(0.4)	1,990	(55.7)	1,507	(42.2)	77	(2.2)	14	(0.4)	~	~	10	(71.4)	~	~
Vietnamese	2,376	(0.3)	1,495	(62.9)	843	(35.5)	38	(1.6)	5	(0.1)	0	(0.0)	~	~	~	~
Other Asian (single group)	16,106	(1.7)	10,556	(65.5)	4,895	(30.4)	655	(4.1)	34	(1.0)	9	(26.5)	21	(61.8)	4	(11.8)

	Male Cohort								Male Incident Lung Cancer Cases							
Characteristic	Total (N=935,638)		Smoking Status						Total (N=3,407)		Smoking Status					
			Never (n=545,543)		Ever (n=350,708)		Unknown (n=39,387)				Never (n=429)		Ever (n=2,463)		Unknown (n=515)	
	N	(%)[d]	N	(%)	N	(%)	N	(%)	N	(%)[d]	N	(%)	N	(%)	N	(%)
Asian (multiple group)	20,506	(2.2)	12,315	(60.1)	7,847	(38.3)	344	(1.7)	100	(2.9)	18	(18.0)	64	(64.0)	18	(18.0)
Asian only multiple	5,113	(0.5)	3,227	(63.1)	1,778	(34.8)	108	(2.1)	8	(0.2)	0	(0.0)	7	(87.5)	~	~
Asian & non-Asian	15,393	(1.6)	9,088	(59.0)	6,069	(39.4)	236	(1.5)	92	(2.7)	18	(19.6)	57	(62.0)	17	(18.5)
Non-Hispanic White	383,466	(41.0)	216,568	(56.5)	153,524	(40.0)	13,374	(3.5)	1,125	(33.0)	127	(11.3)	855	(76.0)	143	(12.7)
Black	21,998	(2.4)	12,119	(55.1)	9,111	(41.4)	768	(3.5)	69	(2.0)	13	(18.8)	53	(76.8)	~	~
Hispanic	68,822	(7.4)	41,585	(60.4)	24,631	(35.8)	2,606	(3.8)	79	(2.3)	17	(21.5)	55	(69.6)	7	(8.9)
Non-AANHPI multiple	24,048	(2.6)	13,224	(55.0)	10,186	(42.4)	638	(2.7)	139	(4.1)	15	(10.8)	102	(73.4)	22	(15.8)
Other (incl AIAN)	18,616	(2.0)	10,861	(58.3)	7,063	(37.9)	692	(3.7)	41	(1.2)	6	(14.6)	29	(70.7)	6	(14.6)
Unknown	235,176	(25.1)	135,613	(57.7)	81,595	(34.7)	17,968	(7.6)	1,339	(39.3)	154	(11.5)	955	(71.3)	230	(17.2)

[a] . Age at baseline among cohort, age at diagnosis among cases.

[b] . Individuals indicating *any* Native Hawaiian, even if also indicating other races/ethnicities, are categorized as ‘Native Hawaiian’.

[c] . Pacific Islander, not indicating Native Hawaiian.

[d] . Column percentages are provided in ‘Total’ columns. Percentages totaling to 100% in ‘Total’ column are mutually exclusive categories of Any AANHPI, Non-Hispanic White, Black, Hispanic, Non-AANHPI multiple, Other (incl AIAN), and Unknown. All other columns with proportions present row percentages.

[e] . The symbol ‘~’ indicates censoring due to low numbers (<5 individuals).

[f] . AANHPI, Asian American, Native Hawaiian, and Pacific Islander; NHPI, Native Hawaiian and Pacific Islander; AIAN, American Indian and Alaska Native.