Published in final edited form as:

Cancer Epidemiol Biomarkers Prev. 2020 April ; 29(4): 796–806. doi:10.1158/1055-9965.EPI-19-0882.

Population-based registry linkages to improve validity of electronic health record-based cancer research

Caroline Thompson[1,2,3], **Anqi Jin**[2], **Harold S. Luft**[2], **Daphne Y. Lichtensztajn**[4,5], **Laura Allen**[4,5], **Su-Ying Liang**[2], **Benjamin Schmacher**[1,3], **Scarlett Lin Gomez**[4,5,6]

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

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

[3]University of California San Diego School of Medicine, San Diego, CA

[4]Greater Bay Area Cancer Registry, Department of Epidemiology & Biostatistics, University of California San Francisco School of Medicine, San Francisco, CA

[5]Department of Epidemiology & Biostatistics, University of California San Francisco School of Medicine, San Francisco, CA

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

Abstract

Background: There is tremendous potential to leverage the value gained from integrating electronic health records (EHRs) and population-based cancer registry data for research. Registries provide detailed diagnosis tumor characteristics, and treatment summaries, while EHRs contain rich clinical detail. A carefully conducted cancer registry linkage may also be used to improve the internal and external validity of inferences made from EHR-based studies.

Methods: We linked the EHRs of a large, multispecialty, mixed-payer healthcare system with the statewide cancer registry and assessed the validity of our linked population. For internal validity, we identify patients that might be “missed” in a linkage, threatening the internal validity of an EHR study population. For generalizability, we compared linked cases with all other cancer patients in the 22-county EHR catchment region.

Results: From an EHR population of 4.5M, we identified 306,554 cancer patients, 26% of the catchment regions cancer patients. 22.7% of linked patients were diagnosed with cancer *after* they migrated away from our healthcare system highlighting an advantage of system-wide linkage. We observed demographic differences between EHR patients and non-EHR patients in the surrounding region and demonstrated use of selection probabilities with model-based standardization to improve generalizability.

Correspondence to: Caroline A. Thompson, San Diego State University, School of Public Health, 5500 Campanile Dr., San Diego, CA 92182, caroline.thompson@sdsu.edu.

Conflict(s) of interest: The authors declare no potential conflicts of interest.

Target issue: Special issue Focus on Modernizing Population Sciences

Conclusions: Our experiences set the foundation to encourage and inform researchers interested in working with EHRs for cancer research as well as provide context for leveraging linkages to assess and improve validity and generalizability.

Impact: Researchers conducting linkages may benefit from considering one or more of these approaches to establish and evaluate the validity of their EHR-based populations.

Keywords

linkage; meaningful use; electronic health records; external validity; causal inference

INTRODUCTION

Cancer research using electronic health records (EHRs) may provide important advantages over the use of cancer registries alone for cancer population health research (1–4). EHRs collect longitudinal data related to health behaviors (e.g., BMI, current smoking status), and preventive care services(5). Appropriate use of EHRs for research can facilitate development of longitudinal studies of environmental or behavioral risk factors, or cancer outcomes after routine screening(6–12). However, EHRs are often hampered by the lack of definitive cancer details in coded fields(13,14). Using only EHRs, accurate identification of cancer cases can be difficult and important characteristics needed to describe a cancer population are often in scanned documents or freetext notes(15,16). A cancer registry linkage, which uses identifying information to match EHR patients with the registry, is a solution for researchers to identify cancer patients and obtain their definitive tumor characteristics(17). Population-based cancer registries are mandated by federal and state law(18), and collect data uniformly on a defined catchment population, while EHRs only collect data reflecting patient care and billing. Registries consolidate information for a given case from multiple sources, and the important data items are also frequently cleaned, adjudicated, and carefully prepared for surveillance activities(19,20).

Registry linkages can be costly and time consuming, and the mechanics of the linkage are often designed with both in mind. “Targeted” linkages begin with an EHR data mining step, designed to identify any “potential” cancer patients by searching for relevant codes (i.e., cancer-specific ICD or CPT codes) and creating a list of suspected cancer patients for linkage(21). Such a targeted search strategy is sufficient for many types of research, and it may be especially reliable for identifying cancers in a single-payer healthcare provider or when patient migration for cancer treatment is unlikely. In a provider environment in which patients are free to seek care across multiple health systems, a targeted linkage may be insufficient. For example, in a study designed to evaluate the relationship between routine cancer screening and downstream cancer outcomes, a patient migrating to a new system between the screening and the cancer diagnosis could be overlooked by a targeted linkage; this patient would be misclassified as cancer-free in an analysis. An alternative to a targeted linkage is a system-wide linkage, in which all EHR patients are matched with the registry, regardless of evidence of cancer in their EHRs, would result in the ascertainment of most, if not all of cancer cases for an entire healthcare system. The system-wide approach additionally serves as a rigorous method for identifying confirmed non-cancer cases – i.e., for selection of controls in case-control studies.

EHRs are an example of “real world data” (i.e., observational healthcare data initially collected for non-research purposes (22–24)) and research use of such data risks numerous potential threats to validity (25–31). Two such threats are: 1) bias due to systematic exclusion of eligible subjects in an EHR population, and 2) bias due to limited generalizability of the EHR to the source population. Broadly, both of these can be viewed as possible sources of selection bias(32), but a key distinction is one of internal vs. external validity. Systematic exclusion of eligible patients in an EHR population is a threat to *internal* validity, i.e., the inference made from the study results may deviate from the true relationship in the EHR population. Inability to generalize the study results to a target population due to nonrandom sampling is a threat to *external* validity, i.e., even if the association derived from the EHR population is internally valid, the inference will not be generalizable to the source population because EHR-based populations in the US healthcare system are usually a convenience sample of persons who happen to go to a specific place for healthcare. When known, sampling fractions are routinely used to improve the generalizability of a study findings(33). For example, if researchers know that their study population differs from the source population by a specific demographic characteristic, they may standardize their results to the demographic distribution(32). If the selection factors are related to a vector of characteristics, model-based standardization may be used to re-weight analyses to the multivariate covariate distribution of the source population(32,34).

In this paper, we describe a system-wide cancer registry linkage undertaken for the adult patient population of a large multispecialty, mixed payer healthcare delivery system located in Northern California. Using a validation sample, we demonstrate what was gained/lost using the system-wide approach over a targeted linkage as a check of internal validity of this approach. We also evaluate the external validity of the linked cancer cohort by comparing it to the surrounding catchment region. Finally, we demonstrate the use of model-based standardization to adjust for improved generalizability.

MATERIALS AND METHODS

EHR Study Population

Sutter Health is a large multispecialty healthcare system serving 22 northern California counties, with more than 4 million patients and 10 million outpatient visits per year. The patient population is diverse: 10% Hispanic, 19% Asian American, Native Hawaiian and Pacific Islander (AANHPI), 21% Black, and a payer mix of 42% PPO, 30% HMO, 23% Medicare/Medicaid, 3% self-payers, and 2% other payers. The EpicCare EHR system, (Epic Systems Corporation, Verona, WI), is used 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 additional transactional detail about care utilization (dates and times, providers seen, etc.).

California Cancer Registry

The California Cancer Registry (CCR) is the National Cancer Institute (NCI)-funded Surveillance Epidemiology and End Results (SEER) program statewide cancer registry. The

CCR monitors the occurrence of all types of cancer (excluding non-melanoma skin cancers) in California, including both new diagnoses and deaths. The CCR includes detailed demographic information, tumor characteristics, and specific details of the first course of treatment for all individual cancer cases occurring in California since 1988. The CCR also geocodes all cancer patients based on home address at time of cancer diagnosis and ascertains follow-up information for long-term survival. The CCR includes data on 5.8M cancer cases overall, adding 190,000 new cases per year.

System-wide Registry Linkage

For the system-wide EHR-CCR linkage, we compiled two "finder files" for comparison. From the EHR, we extracted identifying information for all unique adult (≥18 years old) patients in the EHR (regardless of cancer history). From the CCR, we included all unique individuals diagnosed with cancer in the state of California between 1988 and 2013 (based on the availability of data in the CCR at the time of the study activities).

We used LinkPlus software to conduct a probabilistic linkage of the finder files based on patient name, last 4 digits of the social security number, birthdate, and sex. LinkPlus returns a score, based on the probability of the linkage being a match. Based on a test run of 15,700 randomly sampled matches, we selected a cutoff score of 21.5 and above as a match, which corresponded with a probability of 99.6%. We manually reviewed matches to determine match status for scores between 21.1 and 21.4 (true match % between 65.1% and 80.4%). Scores below 21.1 were not considered as matches. After linkage, we retained three files for further analyses: 1) the EHR cancer patients – i.e., the linked patients and their tumor details, 2) the EHR cancer-free patients, and 3) the non-EHR cancer patients – i.e., the demographic and tumor details for all non-linked cancer patients who were residing in the 22-county Sutter Health catchment region of Northern California at the time of their diagnosis; the third group was explicitly obtained for establishing the external validity of our the EHR cancer population (Figure 1.)

Validation study

A validation study was undertaken to assess: 1) the added benefit of the system-wide (vs. a targeted) linkage approach for identifying cancer patients ("internal validity") and 2) the generalizability of the EHR cancer cohort established with the linkage ("external validity"). For the validation study, we included only first tumors (excluding secondary primary tumors) of the lung/bronchus, colon/rectum, female breast, and prostate in the catchment region diagnosed in 2012 or 2013, (Figure 2) with the sites chosen based on more cancers of higher prevalence and years chosen based on availability of EHR data at all five medical foundations.

For the internal validity study, we used the EHR-matched portion of the validation study and searched all available date ranges in the following EHR tables: medical history, problem list, charges, encounters, medication orders, and laboratory orders for evidence of ICD9 codes that pertain to malignant tumors of the lung/bronchus (162.0–162.9), colon/rectum (153.0–153.9, 154.0–154.1), female breast (174.0–174.9), prostate (185), or unspecified site (198.81–198.82, 198.89, 199.0–199.1), carcinoma in situ (230.3–230.4, 233.0–233.4,

234.9), or history of cancer (V10.00, V10.06, V10.09, V10.11, V10.3, V10.6). This allowed us to identify two subsets of the EHR cancer patients: 1) those with evidence of cancer in their EHR, and 2) those with no evidence of cancer in their EHR. The latter group we presume would not have been identified in a targeted linkage. We compare these groups' demographic and tumor characteristics. For EHR cancer patients who did not have evidence of cancer in their EHRs; we further stratified the population by the timing of their EHR encounters in relationship to their CCR-provided cancer diagnosis date, and evidence of a primary care provider (PCP) assignment. We used the following CCR-provided variables for comparing between the three groups: patient sex, age group (18–54, 55–64, 65–74, 75–84, 85+), race/ethnicity (non-Hispanic White, non-Hispanic Black, Hispanic, Asian American/Pacific Islander, Native American), quintiles of neighborhood socioeconomic status (nSES) derived at the block-group level (35), cancer stage (in situ, localized, regional or remote), primary patient insurance (private, Medicare, any public/Medicaid/military), and marital status (married/registered, single/divorced/widowed).

For the external validity study, we re-defined the EHR cancer patient population to include only patients who had EHR evidence of care during their cancer episode (defined as 90 days prior to and up to 365 days after the CCR-provided date of cancer diagnosis). EHR cancer patients who did not have any care during this timeframe were reclassified as non-EHR cancer patients (Figure 2). By additionally including all non-linked catchment region cancer patients in the non-EHR cancer patients, the total external validation sample is thus equivalent to the underlying source population (cancer patients diagnosed with first primary sites of interest, in 2012–2013, living in 22 Northern CA counties). We compared the two populations (EHR and non-EHR) by county, and according to the characteristics described above, and we calculated ratios of proportion of EHR to non-EHR patients by category, with 95% confidence intervals.

Model-Based Standardization

To demonstrate the use of model-based standardization to re-weight the EHR population based on the covariate distribution of the source population, we used inverse probability of selection weighting (IPSW) to adjust model-based estimates for the relationship between later stage at diagnosis (defined as regional/remote vs in situ/localized) and four exposure variables: nSES quintile, patient race/ethnicity, marital status (married vs. not married) and any public insurance type (vs. private insurance or Medicare). For this demonstration, we excluded unknown values for covariates and patients with very rare covariate values. The assumed data generating mechanism for the relationships modeled, including selection into the EHR population are depicted in a directed acyclic graph (Figure 3). DAGs are nonparametric probabilistic diagrams that depict presumed causal relationships and can be to identify "biasing pathways" that inhibit valid causal inference and to select variables for bias control (36). In the DAG, bias occurs because the selection node, which is predicted by the exposure, outcome, and other covariates in the causal model, is a "collider variable", and conditioning on a collider creates "collider stratification bias" (37). This type of bias can be mitigated by adjusting for any covariates that also predict selection, but if the exposure and outcome also predict selection, bias may occur (38). If it is possible to quantify the selection mechanism, i.e., through a validation study, the pathways may be blocked by re-weighting

the outcome model in a procedure called inverse probability of selection weighting (IPSW) (39,40).

The proof and assumptions required for IPSW have been described elsewhere (34,39). Briefly, let X be our exposure of interest, Y our binary outcome, S a binary selection node (where S=1 for the EHR population and S=0 for non-EHR population), **Z** a vector of confounding variables that are common causes of X, Y, and S, and C is any additional variables that are useful for predicting S but do not confound the XY relationship. In our study c was chosen as the county-specific prevalence of EHR patients (see Supplementary Table 1). We began by modeling the conditional probability of selection $P(S = 1 | y, x, \boldsymbol{z}, c)$ as a logistic regression model including all 2-way product terms:

$$logit\,(P(S = 1 | y, x, \boldsymbol{z}, c)) = \beta_S + \beta_{SY} y + \beta_{SX} x + \beta_{SZ} z + \beta_{SC} c + \beta_{SYX} yx + \beta_{SYZ} yz + \beta_{SYC} yc + \beta_{SXZ} xz + \beta_{SXC} xc$$

This model was run for all study participants and the outputted predicted probabilities were used to calculate individual weights with the marginal probability of the exposure of interest used as the numerator of the stabilized weight (sw), where:

$$sw = P(S = 1 | x) / P(S = 1 | y, x, \boldsymbol{z}, c)$$

For each exposure of interest, we then compared the outcome model parameter estimates calculated three ways: in 1) the entire sample, 2) the EHR sample only, and 3) the EHR sample re-weighted by *sw*. All outcome models were adjusted for sex and continuous age. Additional covariates were included in order to close all biasing pathways, while avoiding adjustment for intermediate variables (32). For example, nSES was included as a covariate in the model for the relationship between marital status and late stage at diagnosis, but not in the model for race/ethnicity and stage, since nSES is an intermediate on the pathway from race/ethnicity to stage. All descriptive and analytical statistics were generated with SAS Enterprise Guide version 10.1 (SAS Institute, Cary, NC).

Results

System-Wide Linkage

Our linkage “finder files” included N=4,816,898 unique adult EHR patients and N=3,350,288 unique CCR patients. The linkage identified a total of N=306,554 Sutter patient matches (group 1), with N=169 likely being duplicate (because multiple EHR patients matched the same CCR patient ID). There were N=840,974 CCR patients residing in the catchment region (group 2) who were not in the Sutter population and N=4,510,344 Sutter patients did not match to the CCR (group 3) (Figure 1). After the linkage, tumor and demographic characteristics for groups 1 and 3 were obtained for a total of 1,338,114 tumors for 1,147,528 unique patients.

The validation study sample (Figure 2) included (N=41,165) patients who were diagnosed with first tumors of the lung/bronchus (N=7,743), colon/rectum (N=6,781), female breast (N=15,953) or prostate (N=10,688) in 2012 or 2013 and residing in the 22-county catchment

region. Initial linkage identified 16,257 as members of the EHR population; this number was reduced after we re-classified the EHR population to include only patients with EHR evidence during their cancer episode (N=10,659), and we compared them to the non-EHR population (N=30,506).

Internal Validity

We found that 12,280 patients (75.5%) had evidence of a past cancer diagnosis in their EHR records, or were being treated at Sutter concurrent to their cancer episode. One thousand three hundred fifty-five (34.08%) of these patients were never assigned a PCP and 3,684 (22.7%) patients did not have evidence of cancer in their EHRs but their cancer diagnosis date was more than 365 days after their last EHR visit. We presume these to be former EHR patients who migrated to another healthcare facility before their cancer diagnosis, or one-time/occasional EHR patients who visited a Sutter hospital or specialist for an orthopedic surgery or delivery but never elected a PCP. Former/occasional patients were more evenly distributed across SES quintiles than current patients. Two-hundred ninety-three (1.8%) patients had no evidence of past cancer diagnosis (>365 days before their Sutter encounters) in their EHRs. These patients were more likely to be higher SES and have been diagnosed with cancers of lower stage (Table 1).

External Validity and Re-Weighting

The EHR population comprised 25.89% of all first cancers diagnosed in this time period and geographic region (Table 2). This proportion ranged by cancer site, from 21.82% for prostate to 29.66% for female breast, and by county, 5.30% in Napa to 64.04% in Yuba (Supplementary Table 1). (The wide range across counties reflects differences in the location of Sutter facilities.) Compared to the non-EHR patients in the catchment region, the EHR population was younger (18–54 PR: 1.12; 95% CI: 1.01–1.25) and less likely to be male (PR: 0.82; 95% CI: 0.71–0.96), non-Hispanic Black (PR: 0.78; 95% CI 0.75–0.82), Hispanic (PR: 0.74; 95% CI: 0.72–0.77), or Asian American/Pacific Islander (PR: 0.78; 95% CI: 0.72–0.86), more likely to be higher SES (highest nSES PR: 1.17; 95% CI: 1.03–1.33; lowest nSES PR: 0.93; 95% CI: 0.91–0.96). For insurance type, we observed that a larger proportion of EHR patients claimed Medicare as their primary payer (PR: 1.55; 95% CI: 1.37–1.74), and less used public insurance (PR: 0.71; 95% CI: 0.69–0.73). These patterns varied little by tumor site, with a few exceptions. EHR lung cancer patients were slightly overrepresented non-Hispanic Blacks (PR: 1.08; 95% CI: 1.02–1.14) and their SES distribution was more similar to that of the underlying population. For tumor stage, EHR patients had more in situ cancers overall (PR: 1.22; 95% CI: 1.15–1.28) but these differences disappeared when stratified by site (Supplementary Table 2).

For the IPSW demonstration, we excluded 6,849 patients (16.6%) with unknown and very rare covariate values, resulting in a final analytical sample of 34,316. For model 1, we observed positive relationship between lower SES and later stage, which appeared to follow a linear trend (OR for lowest nSES: 1.958; 95% CI: 1.801–2.129). The unweighted EHR-only models slightly exaggerated these relationships (OR for lowest nSES: 2.022; 95% CI: 1.719–2.378). For model 2, being unmarried was associated with an increased odds of later stage at diagnosis in the full sample (OR: 1.394; 95% CI: 1.329–1.462) and this relationship

was slightly attenuated in the unweighted EHR population. For model 3, in the full sample, having public insurance was strongly associated with later stage at diagnosis (OR: 1.901; 95% CI: 1.757–2.055) and also slightly attenuated in the EHR-only population (OR: 1.844; 95% CI: 1.553–2.190). Finally, compared to non-Hispanic Whites, non-Hispanic Blacks (OR: 1.135; 95% CI: 1.043–1.234) and Hispanics (OR: 1.184; 95% CI: 1.104–1.269) had higher adjusted odds of later stage at diagnosis, and these relationships were slightly exaggerated in the EHR-only analysis. The unweighted EHR odds ratio for Asian American and Pacific Islanders was higher, but not significantly different from the null, but in the catchment region and in the re-weighted EHR population, AAPIs were more likely to be diagnosed at later stage compared to non-Hispanic Whites (OR: 1.164; 95% CI: 1.028–1.307). For all models, the IPSW procedure was effective at adjusting the odds ratios in EHR population so that they more closely resembled the full catchment region (Table 3).

DISCUSSION

We were able to link a large healthcare system with a statewide population-based cancer registry in order to identify cancer patients. Our validation study was designed to demonstrate improved cancer case ascertainment with a system-wide linkage approach and to evaluate the representativeness of our resulting EHR-based cancer cohort, by comparing it to the underlying catchment region.

For our internal validity study, we identified 22.7% of all linkage-identified EHR cancer patients who were diagnosed with cancer *after* they migrated away from our healthcare system and a 1.8% of patients who had a history of cancer that was not recorded in their EHRs. The accuracy/representativeness of evidence-in-EHR vs. not-in-EHR in our study may demonstrate an “inverse survivorship bias”, i.e., patients with more treatable cancers (e.g., colon/rectum) leading to longer survival and hence, more chance of system migration, while less treatable cancers (e.g., lung) have shorter survival and thus less time for migration. While the pattern did not hold for breast cancers (most of which are treatable) this is a type of selection bias(38) and should be considered when relying on EHRs to study cancers of differing prognoses.

The proportion of registry-identified cancers that were not represented in the EHRs points to weaknesses in the targeted linkage approach, however, the implications of these findings depend on the study design. A cohort study of risk factors for cancer based on longitudinal follow-up would suffer from substantial bias if the patients who were diagnosed with cancer after migration remained classified as cancer-free, given that 23% of patients were subsequently identified with cancer via linkage to a population-based cancer registry. In contrary, a study of patients treated for their cancer at the healthcare system would likely be unharmed by this omission, given only 2% of cases omitted a history of cancer in their EHR. We also found that 34% of Sutter cancer patients who were not represented in the EHR were never assigned a PCP, so the nature of their affiliation to the EHR system was questionable to begin with, and so the study with a narrower focus on just those patients who also had a PCP would be quite robust. A third approach to generating finder files for linkage might be to start with a subset of the patients of interest, for whom key variables might be expected to be collected, e.g., the primary care base or by selecting a subset of patients based on

information density, which has been identified as an important potential indicator of EHR data quality(41).

Self-selection of patients to a particular healthcare system is a complex multifactorial mechanism. With some notable exceptions, the EHR population in our study was generally representative of the underlying catchment region. We observed some demographic differences between EHR patients and non-EHR patients in the surrounding region, which may be partially explained by characteristics of the region or healthcare system. For example, EHR patients were more often female, and older age EHR patients claimed Medicare coverage more often. Some possible explanations for these findings include the availability of Sutter breast cancer specialists in some regions, and presence of large competitor HMO systems. These findings highlight that knowledge of provider availability and market characteristics of catchment region are important for interpretation of these results, and generally for research use of EHR data.

We demonstrated the use of selection probabilities with model-based standardization to improve generalizability of our EHR population to the underlying catchment region. We did not observe substantial differences in the conditional odds of our outcome (late stage at diagnosis) between the EHR population and the full catchment region. Indeed, the re-weighting procedure resulted in only one odds ratio that would have changed our statistical inference compared to the unweighted results. This could be an additional indication of the generalizability of our EHR cancer patient population. Alternatively, the observed differences in demographic distributions may not be important for the modeled outcome across strata of the selected and non-selected populations. Either way, undertaking a simple comparison of modeled results between the EHR and the catchment region (even in the absence of implementing IPSW) serves to strengthen an EHR study's external validity.

In order for IPSW to be valid for identifying the causal effect, some strong assumptions are required: there must be absence of other systematic error (confounding, measurement error), and the specified model for probability of selection must be sufficient for rendering the non-selected missing at random. Our IPSW demonstration was fairly ideal in that we had access to the all covariates in both the selected and non-selected population. In a more common scenario, one or more of the variables (e.g., EHR-derived) may predict selection and be unavailable for non-EHR patients. These important unmeasured predictors of selection, such as availability of employer-based healthcare coverage, or health literacy, cannot be overcome by IPSW, and the credibility of this approach relies on a realistic scenario and robust causal diagram. A less model-driven approach is also possible with knowledge of some but not all selection probabilities, which can be used to adjust descriptive statistics derived from an EHR population, such as incidence or prevalence rates, as is done in weighted survey design.

We found just a few examples of studies that were similar to ours in objective and scope. Related to our findings from the internal validity study, Clarke et al. demonstrated the added value of EHRs to identify patients with a history of cancer who might not appear in the statewide tumor registry(15). Relevant to our external validation study, there have been two studies from Kaiser that aimed to characterize the generalizability of their EHR populations, in breast(42), and lung(43) cancer patients. Selection bias is a known concern for EHR-

based research. Hanuese and Daniels developed a framework for evaluating such bias by comparing subsets of a patient population with(out) EHR-derived covariate information(44). Their framework (and other similar studies(45–48)), have emphasized the importance of data provenance (i.e., understanding the technology- and provider-related factors that impact how and why EHR data are generated) when considering bias in EHR research. Based on our findings, geography and regional context could be additional candidates for important data provenance considerations(49). IPSW has been used for the purposes of generalizing randomized controlled trial data(50) and generalizing autopsy data to a live source population(51). We also found one instance of its use with EHRs in a study of childhood obesity(52).

EHR systems are vast databases that do not have easily accessible research-ready data tables. Research use of EHRs requires a knowledgeable support team experienced in interpreting researcher questions, and extracting the necessary data. Large scale data initiatives like the one we have described also rely on sharing protected health information (PHI) across institutions in order to improve scale and validity, but maintaining patient privacy is a key challenge. This requires both secure processes and multiple organizational agreements. We obtained all necessary privacy and legal approvals (from all organizations involved) and extracted names birthdates, sex, zip codes, and last 4 digits of social security number for 4.5 million adult members of an EHR population. Due to the size of the populations studied, consent was not sought for participation in this study, but we instead obtained authorization for a waiver of consent. Upon return of the cancer details for successfully linked patients (with personal identifiers removed), we additionally extracted “limited” (excluding direct personal identifiers) EHR data pertaining to select patients’ cancer care for the validation study. New approaches may allow HIPAA-covered entities to share data for patient identification and linkage across data sourcesand greatly reduce the time and effort required to accomplish a study like ours(53).

Limitations

Linkage with the statewide registry only ascertains reportable cases, i.e., those who lived in the catchment area at the time of diagnosis. Out of state residents who sought care at a Sutter facility would not be captured. False positives are possible with probabilistic linkage, however it has been shown to be valid and, compared to deterministic linkages processes, it is better suited for large data (54–56). Our choice of cancers in the validation sample may impact the results of our internal validity checks. For example, some system providers may be well respected in the medical community and be regularly sought out only for second opinions, which would increase the number of cancer patients who link to the statewide registry, but who are receiving the majority of their care elsewhere..

Conclusions

EHR-based study populations are a convenience sample, but the efficiency of such studies often outweighs the drawbacks. The representativeness of any research database has important implications for the generalizability of findings and recommendations for policy or evidenced-based treatment strategies. Our experiences help encourage and inform

researchers interested in working with EHRs for cancer research as well as provide context for leveraging linkages to assess and improve study validity and generalizability.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

The authors would like to extend their thanks and gratitude to the research staff who supported this effort, including Rita Leung, Sarah Knowles, Pragati Kenkare, and Mai Vu. Funding for the linkage was provided by the National Cancer Institute as part of a Surveillance Epidemiology and End Results (SEER) Rapid Response Surveillance Study on Patient Generated Health Data (HHSN261201300005I). CT was funded by a career development award from the National Institute for Advancing Translational Sciences (KL2TR001444). The collection of cancer incidence data used in this study was supported by the California Department of Public Health pursuant to California Health and Safety Code Section 103885; Centers for Disease Control and Prevention's (CDC) National Program of Cancer Registries, under cooperative agreement 5NU58DP006344; the National Cancer Institute's Surveillance, Epidemiology and End Results Program under contract HHSN261201800032I awarded to the University of California, San Francisco, contract HHSN261201800015I awarded to the University of Southern California, and contract HHSN261201800009I awarded to the Public Health Institute, Cancer Registry of Greater California. The ideas and opinions expressed herein are those of the author(s) and do not necessarily reflect the opinions of the State of California, Department of Public Health, the National Cancer Institute, and the Centers for Disease Control and Prevention or their Contractors and Subcontractors.

REFERENCES

1. Yu P, Artz D, Warner J. Electronic health records (EHRs): supporting ASCO's vision of cancer care. Am Soc Clin Oncol Educ Book 2014:225–31 doi 10.14694/EdBook_AM.2014.34.225. [PubMed: 24857080]
2. Yu PP. The evolution of oncology electronic health records. Cancer J 2011;17(4):197–202 doi 10.1097/PPO.0b013e3182269629. [PubMed: 21799325]
3. Miriovsky BJ, Shulman LN, Abernethy AP. Importance of health information technology, electronic health records, and continuously aggregating data to comparative effectiveness research and learning health care. J Clin Oncol 2012;30(34):4243–8 doi 10.1200/JCO.2012.42.8011. [PubMed: 23071233]
4. Warner J, Hochberg E. Where is the EHR in oncology? J Natl Compr Canc Netw 2012;10(5):584–8. [PubMed: 22570289]
5. Weiner MG, Lyman JA, Murphy S, Weiner M. Electronic health records: high-quality electronic data for higher-quality clinical research. Inform Prim Care 2007;15(2):121–7. [PubMed: 17877874]
6. Hughes AE, Tiro JA, Balasubramanian BA, Skinner CS, Pruitt SL. Social Disadvantage, Healthcare Utilization, and Colorectal Cancer Screening: Leveraging Longitudinal Patient Address and Health Records Data. Cancer Epidemiology Biomarkers & Prevention 2018;27(12):1424–32 doi 10.1158/1055-9965.Epi-18-0446.
7. Thompson CA, Gomez SL, Chan A, Chan JK, McClellan SR, Chung S, et al. Patient and Provider Characteristics Associated with Colorectal, Breast, and Cervical Cancer Screening among Asian Americans. Cancer Epidemiology Biomarkers & Prevention 2014;23(11):2208–17 doi 10.1158/1055-9965.Epi-14-0487.
8. Mayer MA, Gutierrez-Sacristan A, Leis A, De La Pena S, Sanz F, Furlong LI. Using Electronic Health Records to Assess Depression and Cancer Comorbidities. Stud Health Technol Inform 2017;235:236–40. [PubMed: 28423789]
9. Young-Wolff KC, Klebaner D, Folck B, Tan ASL, Fogelberg R, Sarovar V, et al. Documentation of e-cigarette use and associations with smoking from 2012 to 2015 in an integrated healthcare delivery system. Preventive Medicine 2018;109:113–8 doi 10.1016/j.ypmed.2018.01.012. [PubMed: 29360481]

10. Huo J, Yang M, Tina Shih Y-C. Sensitivity of Claims-Based Algorithms to Ascertain Smoking Status More Than Doubled with Meaningful Use. Value in Health 2018;21(3):334–40 doi 10.1016/j.jval.2017.09.002. [PubMed: 29566841]
11. Schinasi LH, Auchincloss AH, Forrest CB, Diez Roux AV. Using electronic health record data for environmental and place based population health research: a systematic review. Annals of Epidemiology 2018;28(7):493–502 doi 10.1016/j.annepidem.2018.03.008. [PubMed: 29628285]
12. Cole AM, Pflugeisen B, Schwartz MR, Miller SC. Cross sectional study to assess the accuracy of electronic health record data to identify patients in need of lung cancer screening. BMC Research Notes 2018;11(1):14 doi 10.1186/s13104-018-3124-0. [PubMed: 29321038]
13. Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. International Journal of Medical Informatics 2008;77(5):291–304 doi 10.1016/j.ijmedinf.2007.09.001. [PubMed: 17951106]
14. Vuokko R, Mäkelä-Bengs P, Hyppönen H, Lindqvist M, Doupi P. Impacts of structuring the electronic health record: Results of a systematic literature review from the perspective of secondary use of patient data. International Journal of Medical Informatics 2017;97:293–303 doi 10.1016/j.ijmedinf.2016.10.004. [PubMed: 27919387]
15. Clarke CL, Feigelson HS. Developing an Algorithm to Identify History of Cancer Using Electronic Medical Records. EGEMS (Washington, DC) 2016;4(1):1209- doi 10.13063/2327-9214.1209.
16. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. JMIR Med Inform 2019;7(2):e12239 doi 10.2196/12239.
17. Jacobs EJ, Briggs PJ, Deka A, Newton CC, Ward KC, Kohler BA, et al. Follow-up of a Large Prospective Cohort in the United States Using Linkage With Multiple State Cancer Registries. American Journal of Epidemiology 2017;186(7):876–84 doi 10.1093/aje/kwx129. [PubMed: 28520845]
18. Cancer Registries Amendment Act. In: (1991-1992) nC, editor. Volume S.33121992.
19. National Program of Cancer Registries Program Standards. Centers for Disease Control and Prevention.
20. Thoburn KK, German RR, Lewis M, Nichols PJ, Ahmed F, Jackson-Thompson J. Case completeness and data accuracy in the Centers for Disease Control and Prevention's National Program of Cancer Registries. Cancer 2007;109(8):1607–16 doi 10.1002/cncr.22566. [PubMed: 17343277]
21. Kurian AW, Mitani A, Desai M, Yu PP, Seto T, Weber SC, et al. Breast cancer treatment across health care systems: linking electronic medical records and state registry data to enable outcomes research. Cancer 2014;120(1):103–11 doi 10.1002/cncr.28395. [PubMed: 24101577]
22. 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(23):2293–7 doi 10.1056/NEJMsb1609216. [PubMed: 27959688]
23. Mahajan R. Real world data: Additional source for making clinical decisions. Int J Appl Basic Med Res 2015;5(2):82 doi 10.4103/2229-516X.157148. [PubMed: 26097811]
24. Khozin S, Blumenthal GM, Pazdur R. Real-world Data for Clinical Evidence Generation in Oncology. J Natl Cancer Inst 2017;109(11) doi 10.1093/jnci/djx187.
25. 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. BMC Med Inform Decis Mak 2014;14:51 doi 10.1186/1472-6947-14-51. [PubMed: 24916006]
26. 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(4):346–52 doi 10.1007/s40471-017-0130-z. [PubMed: 31223556]
27. 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(5):e185 doi 10.2196/jmir.9134. [PubMed: 29844010]
28. 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(11):847–55 doi 10.1093/aje/kww112. [PubMed: 27852603]

29. 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(6):1134–41 doi 10.1093/jamia/ocx071. [PubMed: 29016972]
30. Desai JR, Wu P, Nichols GA, Lieu TA, O’Connor PJ. Diabetes and asthma case identification, validation, and representativeness when using electronic health data to construct registries for comparative effectiveness and epidemiologic research. Med Care 2012;50 Suppl:S30–5 doi 10.1097/MLR.0b013e318259c011. [PubMed: 22692256]
31. Stuart EA, DuGoff E, Abrams M, Salkever D, Steinwachs D. Estimating causal effects in observational studies using Electronic Health Data: Challenges and (some) solutions. EGEMS (Wash DC) 2013;1(3) doi 10.13063/2327-9214.1038.
32. Rothman K, Greenland S, Lash T. Modern Epidemiology, 3rd Edition. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
33. Kalton G, Flores-Cervantes I. Weighting Methods. Journal of Official Statistics 2003;19(2):81–97.
34. Thompson CA, Arah OA. Selection bias modeling using observed data augmented with imputed record-level probabilities. Ann Epidemiol 2014;24(10):747–53 doi 10.1016/j.annepidem.2014.07.014. [PubMed: 25175700]
35. Yang J, Schupp C, Harrati A, Clarke C, Keegan T, Gomez S. Developing an area-based socioeconomic measure from American Community Survey data. Fremont, CA: Cancer Prevention Institute of California; 2014.
36. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology 1999;10(1):37–48. [PubMed: 9888278]
37. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. Epidemiology 2003;14(3):300–6. [PubMed: 12859030]
38. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004;15(5):615–25. [PubMed: 15308962]
39. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000;11(5):550–60. [PubMed: 10955408]
40. Mansournia MA, Altman DG. Inverse probability weighting. BMJ 2016;352:i189 doi 10.1136/bmj.i189.
41. Reimer AP, Milinovich A, Madigan EA. Data quality assessment framework to assess electronic medical record data for use in research. International Journal of Medical Informatics 2016;90:40–7 doi 10.1016/j.ijmedinf.2016.03.006. [PubMed: 27103196]
42. Gomez SL, Shariff-Marco S, Von Behren J, Kwan ML, Kroenke CH, Keegan TH, et al. Representativeness of breast cancer cases in an integrated health care delivery system. BMC Cancer 2015;15:688 doi 10.1186/s12885-015-1696-9. [PubMed: 26467773]
43. Check DK, Albers KB, Uppal KM, Suga JM, Adams AS, Habel LA, et al. Examining the role of access to care: Racial/ethnic differences in receipt of resection for early-stage non-small cell lung cancer among integrated system members and non-members. Lung Cancer 2018;125:51–6 doi 10.1016/j.lungcan.2018.09.006. [PubMed: 30429038]
44. Haneuse S, Daniels M. A General Framework for Considering Selection Bias in EHR-Based Studies: What Data Are Observed and Why? EGEMS (Wash DC) 2016;4(1):1203 doi 10.13063/2327-9214.1203. [PubMed: 27668265]
45. Johnson KE, Kamineni A, Fuller S, Olmstead D, Wernli KJ. How the provenance of electronic health record data matters for research: a case example using system mapping. EGEMS (Wash DC) 2014;2(1):1058 doi 10.13063/2327-9214.1058. [PubMed: 25821838]
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(1):1127 doi 10.13063/2327-9214.1127. [PubMed: 25992389]
47. Hersh WR, Cimino J, Payne PR, Embi P, Logan J, Weiner M, et al. Recommendations for the use of operational electronic health record data in comparative effectiveness research. EGEMS (Wash DC) 2013;1(1):1018 doi 10.13063/2327-9214.1018. [PubMed: 25848563]
48. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care 2013;51(8 Suppl 3):S30–7 doi 10.1097/MLR.0b013e31829b1dbd. [PubMed: 23774517]

49. Kroneman M, Verheij R, Tacken M, van der Zee J. Urban-rural health differences: primary care data and self reported data render different results. Health Place 2010;16(5):893–902 doi 10.1016/j.healthplace.2010.04.015. [PubMed: 20493756]
50. Buchanan AL, Hudgens MG, Cole SR, Mollan KR, Sax PE, Daar ES, et al. Generalizing Evidence from Randomized Trials using Inverse Probability of Sampling Weights. J R Stat Soc Ser A Stat Soc 2018;181(4):1193–209 doi 10.1111/rssa.12357.
51. Haneuse S, Schildcrout J, Crane P, Sonnen J, Breitner J, Larson E. Adjustment for selection bias in observational studies with application to the analysis of autopsy data. Neuroepidemiology 2009;32(3):229–39 doi 10.1159/000197389. [PubMed: 19176974]
52. Flood TL, Zhao YQ, Tomayko EJ, Tandias A, Carrel AL, Hanrahan LP. Electronic health records and community health surveillance of childhood obesity. Am J Prev Med 2015;48(2):234–40 doi 10.1016/j.amepre.2014.10.020. [PubMed: 25599907]
53. PRNewswire. Datavant partners with the People-Centered Research Foundation to de-identify and link data across national clinical research network. San Francisco2019.
54. Clark DE, Hahn DR. Comparison of probabilistic and deterministic record linkage in the development of a statewide trauma registry. Proc Annu Symp Comput Appl Med Care 1995:397–401. [PubMed: 8563310]
55. Tromp M, Ravelli AC, Bonsel GJ, Hasman A, Reitsma JB. Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. J Clin Epidemiol 2011;64(5):565–72 doi 10.1016/j.jclinepi.2010.05.008. [PubMed: 20952162]
56. Garvin JH, Herget KA, Hashibe M, Kirchhoff AC, Hawley CW, Bolton D, et al. Linkage between Utah All Payers Claims Database and Central Cancer Registry. Health Serv Res 2019;54(3):707–13 doi 10.1111/1475-6773.13114. [PubMed: 30675913]

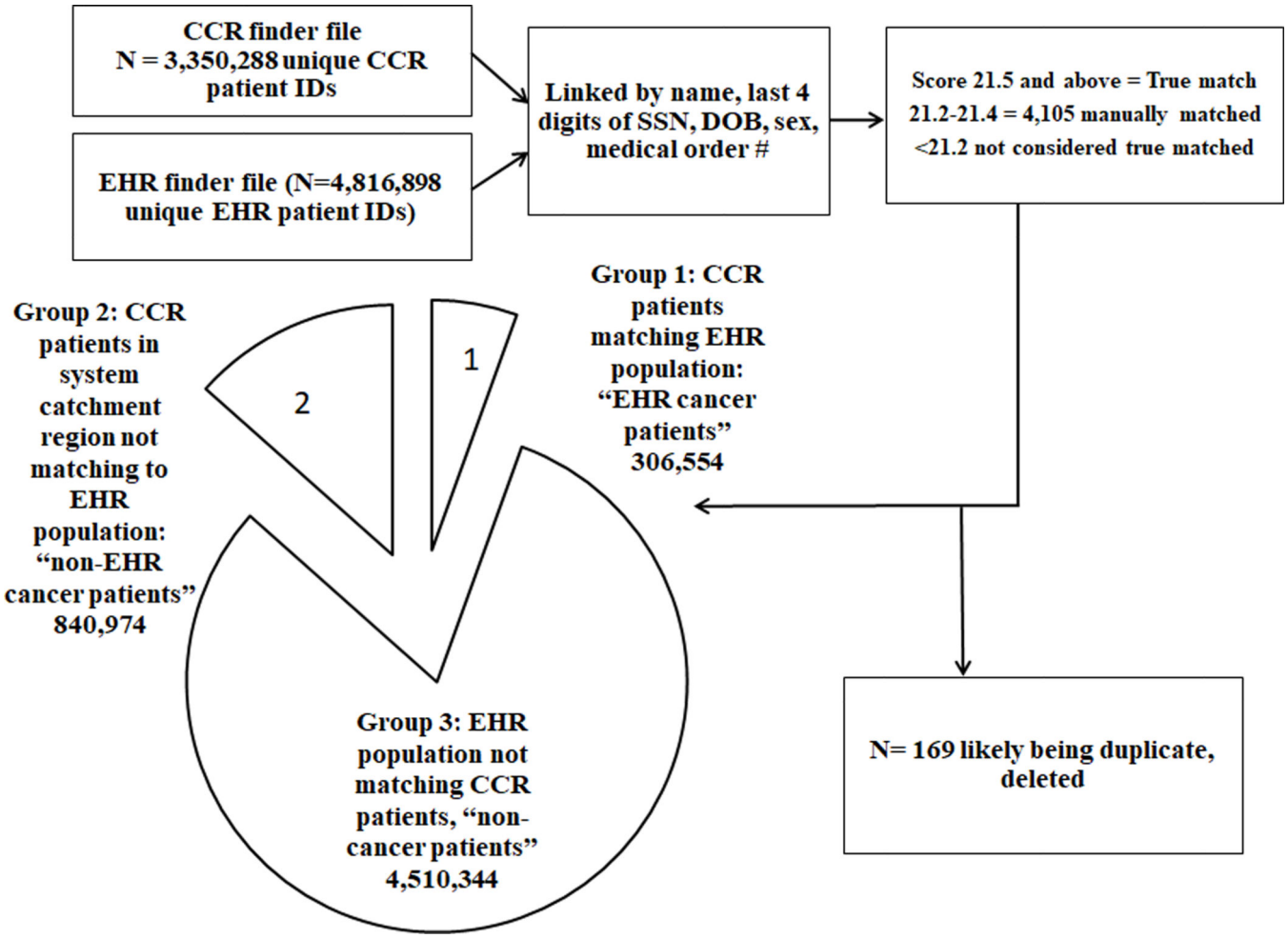

Figure 1:

Flow diagram of the system-wide linkage of ~4.8M EHR patients with ~3.5M CCR patients, yielding three distinct groups: 1) EHR patients matching CCR patients (EHR cancer population); 2) CCR patients not matching EHR patients but residing in the 22-county catchment region (non-EHR cancer population); 3) EHR patients not matching CCR patients (cancer-free EHR patients).

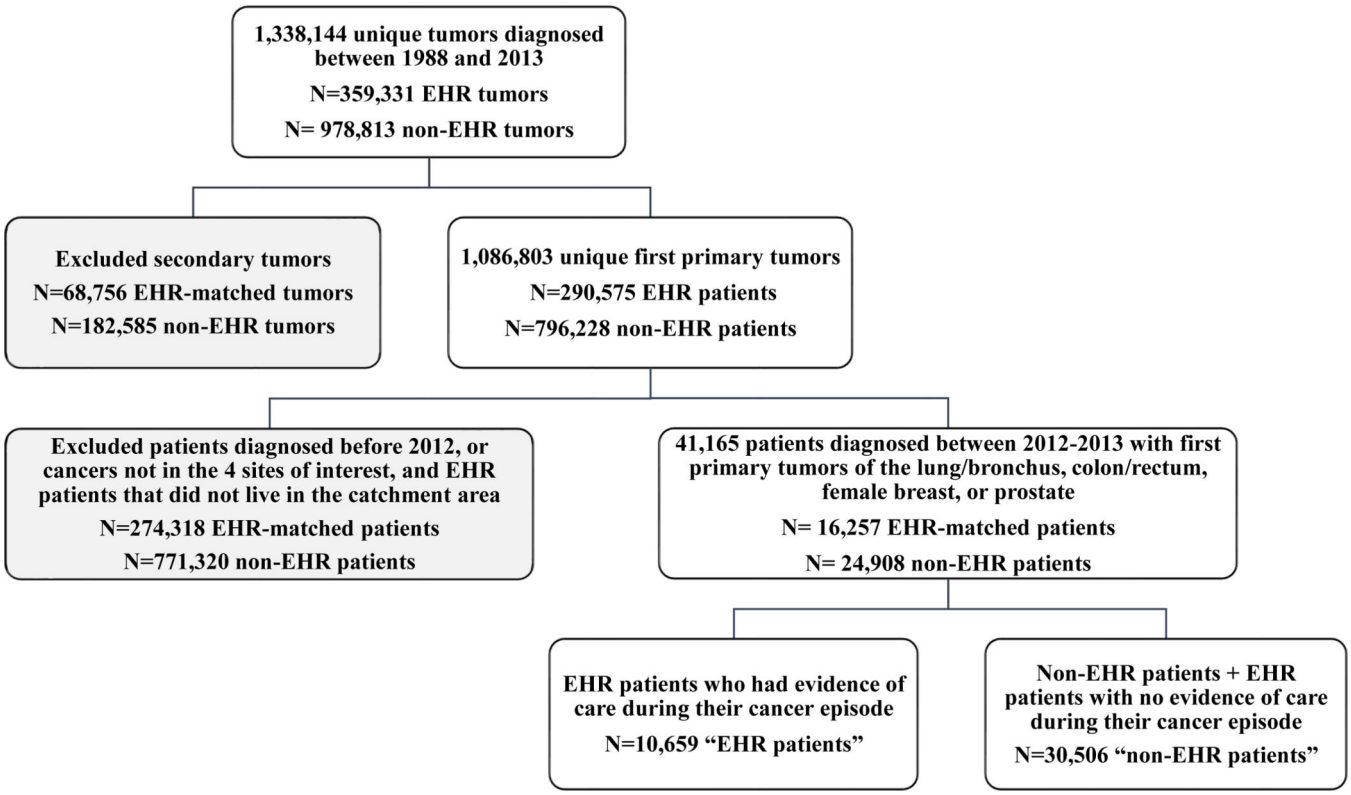

Figure 2:
Definition of the Validation study sample (N=10,659 EHR patients and N=30,506 non-EHR patients)

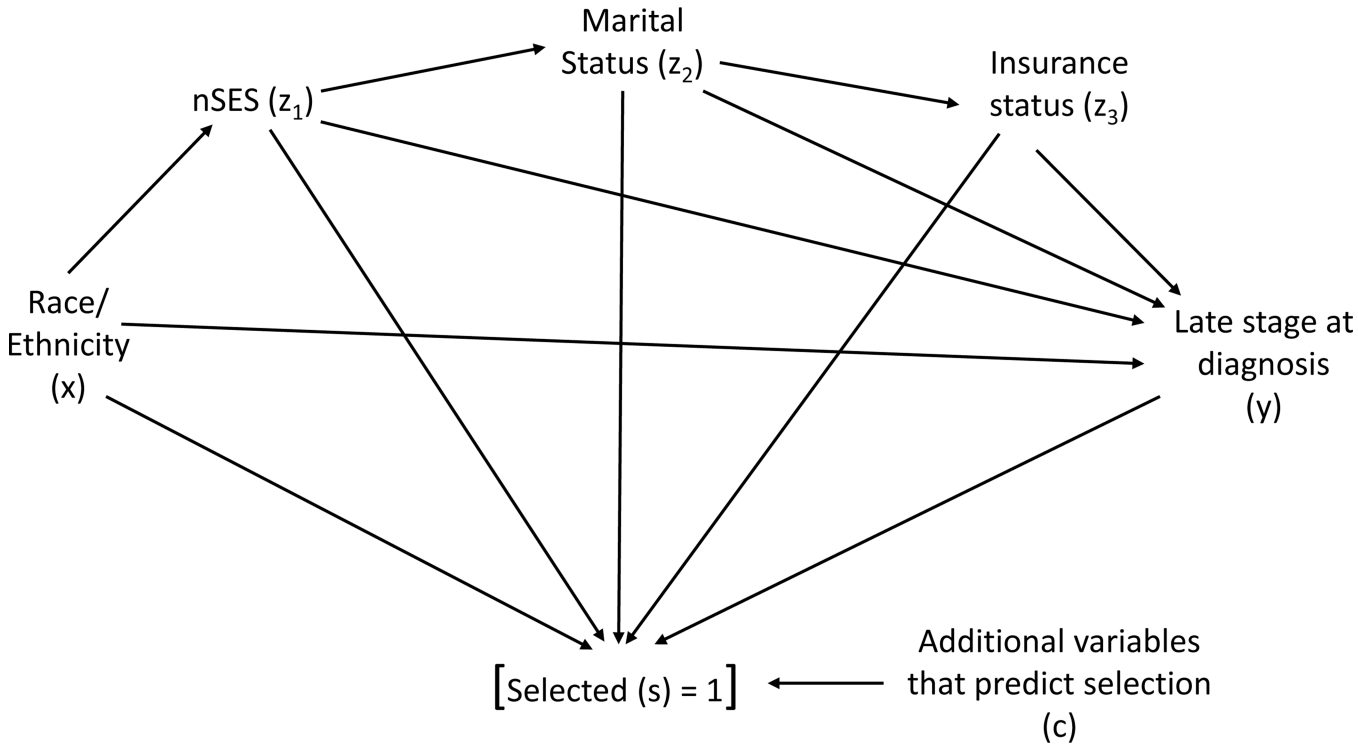

Figure 3:
Directed Acyclic Graph (DAG) depicting the causal relationship and selection mechanism modeled in the inverse probability of selection weighing (IPSW) demonstration, with race/ethnicity depicted as the primary exposure of interest, late stage at diagnosis as the outcome, three confounding variables, conditioned selection node, and predictors of selection.

Table 1.

Internal Validity: EHR Evidence of Cancer Diagnosis in the Registry-Linked EHR Population (N=16,257)

Characteristics of Patients	Evidence of Cancer in EHR (N=12,280)	No Evidence of Cancer In EHR (N=3,977)	No Evidence of Cancer In EHR (N=3,977)		
			EHR Encounters Pre-Date Cancer Diagnosis[a] (N=3,684)	EHR Encounters Post-Date Cancer Diagnosis[b] (N=293)	Primary Care Provider never assigned (N=1,355)
	N (%)	N (%)	N (%)	N (%)	N (%)
Tumor Site					
Lung/Bronchus	2,346 (19.10)	565 (14.21)	558 (15.15)	Low[1]	213 (15.72)
Colon/Rectum	1,716 (13.97)	830 (20.87)	745 (20.22)	85 (29.01)	291 (21.48)
Female Breast	5,697 (46.39)	1,310 (32.94)	1,249 (33.90)	61 (20.82)	382 (28.19)
Prostate	2,521 (20.53)	1,272 (31.98)	1,132 (30.73)	140 (47.78)	469 (34.61)
Sex					
Male	4,407 (35.89)	1915 (48.15)	1,731 (46.99)	184 (62.8)	699 (51.51)
Female	7,873 (64.11)	2062 (51.85)	1,953 (53.01)	109 (37.2)	656 (48.41)
Age Group					
18–54 years	3,150 (25.65)	940 (23.64)	871 (23.64)	69 (23.55)	273 (20.15)
55–64 years	3,182 (25.91)	1153 (28.99)	1,056 (28.66)	97 (33.11)	407 (30.04)
65–74 years	3,364 (27.39)	1073 (26.98)	999 (27.12)	74 (25.26)	379 (27.97)
75–84 years	1,860 (15.15)	585 (14.71)	544 (14.77)	41 (13.99)	214 (15.79)
85 years and older	724 (5.90)	226 (5.68)	214 (5.81)	12 (4.10)	82 (6.05)
Race/ethnicity					
Non-Hispanic White	8,107 (66.02)	2542 (63.92)	2,374 (64.44)	168 (57.34)	864 (63.76)
Non-Hispanic Black	832 (6.78)	295 (7.42)	268 (7.27)	27 (9.22)	118 (8.71)
Hispanic	1,156 (9.41)	413 (10.38)	381 (10.34)	32 (10.92)	133 (9.82)
Asian American/Pacific Islander	1,850 (15.07)	581 (14.61)	534 (14.50)	47 (16.04)	199 (14.69)
Native American	69 (0.56)	14 (0.35)	12 (0.33)	Low[1]	Low[1]
Other/Unknown	266 (2.17)	132 (3.32)	115 (3.12)	17 (5.80)	69 (5.02)
Stage					
In Situ	1,198 (9.76)	303 (7.62)	286 (7.76)	17 (5.80)	71 (5.24)
Localized	5,935 (48.33)	2108 (53.00)	1,920 (52.12)	188 (64.16)	715 (52.77)

Characteristics of Patients	Evidence of Cancer in EHR (N=12,280)	No Evidence of Cancer In EHR (N=3,977)	No Evidence of Cancer In EHR (N=3,977)		
			EHR Encounters Pre-Date Cancer Diagnosis[a] (N=3,684)	EHR Encounters Post-Date Cancer Diagnosis[b] (N=293)	Primary Care Provider never assigned (N=1,355)
	N (%)	N (%)	N (%)	N (%)	N (%)
Regional	2,780 (22.64)	828 (20.82)	767 (20.82)	61 (20.82)	284 (20.96)
Remote	2,119 (17.26)	598 (15.04)	576 (15.64)	22 (7.51)	217 (16.01)
Unknown	248 (2.02)	140 (3.52)	135 (3.66)	Low[1]	68 (5.02)
Neighborhood SES (Quintiles)					
Highest SES	4,021 (32.74)	996 (25.04)	902 (24.48)	94 (32.08)	296 (21.85)
Higher-middle SES	2,951 (24.03)	985 (24.77)	902 (24.48)	83 (28.33)	319 (23.54)
Middle SES	2,419 (19.70)	877 (22.05)	810 (21.99)	67 (22.87)	317 (23.39)
Lower-middle SES	1,763 (14.36)	647 (16.27)	616 (16.72)	31 (10.58)	239 (17.64)
Lowest SES	1,126 (9.17)	472 (11.87)	454 (12.32)	18 (6.14)	184 (13.58)
Payer (Insurance Type)					
Private Insurance	5,669 (46.16)	2,118 (53.26)	1,961 (53.23)	157 (53.58)	628 (46.35)
Medicare	4,649 (37.86)	1,318 (33.14)	1,224 (33.22)	94 (32.08)	482 (35.57)
Any public, Medicaid, military	874 (7.12)	326 (8.20)	306 (8.31)	20 (6.83)	140 (10.33)
Not Insured	100 (0.81)	20 (0.50)	18 (0.49)	Low[1]	13 (0.96)
Unknown	988 (8.05)	195 (4.90)	175 (4.75)	20 (6.83)	92 (6.79)
Marital Status					
Married/Registered	7,143 (58.17)	2,177 (54.74)	2,019 (54.80)	158 (53.92)	721 (53.21)
Separated/Divorced/Widowed	4,558 (37.12)	1,369 (34.42)	1,267 (34.39)	102 (34.81)	491 (36.24)
Unknown	579 (4.71)	431 (10.84)	398 (10.8)	33 (11.26)	143 (10.55)

[a] If the patient's EHR encounters occurred more than 365 days before the cancer diagnosis, we considered them a "former" patient of the healthcare system.

[b] If the patient's EHR encounters occurred more than 365 days after the cancer diagnosis, we considered them to be missing a past history of cancer in their records.

Table 2.

External Validity: Comparison of Demographic and Tumor Characteristics, All Tumor Sites (N=41,165)

Patient Characteristic	EHR Population (N=10,659)	Non-EHR Population (N=30,506)	EHR: Non-EHR
	N (%)	N (%)	Proportion Ratio[a](95% CI[b])
Cancer Site			
Lung/Bronchus	1,993 (18.70)	5,750 (18.85)	0.99 (0.91–1.09)
Colon/Rectum	1,603 (15.04)	5,178 (16.97)	0.89 (0.82–0.96)
Female Breast	4,731 (44.39)	11,222 (36.79)	1.21 (1.05–1.39)
Prostate	2,332 (21.88)	8,356 (27.39)	0.80 (0.71–0.90)
Sex			
Male	3,995 (37.48)	13,882 (45.51)	0.82 (0.71–0.96)
Female	6,664 (62.52)	16,624 (54.49)	1.15 (0.98–1.35)
Age Group			
18–54 years	2,642 (24.79)	6,722 (22.04)	1.12 (1.01–1.25)
55–64 years	2,728 (25.59)	8,586 (28.15)	0.91 (0.80–1.03)
65–74 years	2,941 (27.59)	8,819 (28.91)	0.95 (0.84–1.08)
75–84 years	1,671 (15.68)	4,590 (15.05)	1.04 (0.98–1.11)
85 years and older	677 (6.35)	1,789 (5.86)	1.08 (0.94–1.25)
Race/Ethnicity			
Non-Hispanic White	7,149 (67.07)	17,759 (58.21)	1.15 (0.98–1.36)
Non-Hispanic Black	692 (6.49)	2,530 (8.29)	0.78 (0.75–0.82)
Hispanic	993 (9.32)	3,817 (12.51)	0.74 (0.72–0.77)
Asian American/Pacific Islander	1,533 (14.38)	5,595 (18.34)	0.78 (0.72–0.86)
Native American	56 (0.53)	143 (0.47)	1.12 (0.02–62.43)
Other/Unknown	236 (2.21)	662 (2.17)	1.02 (0.50–2.09)
Stage			
In situ	1,025 (9.62)	2,412 (7.91)	1.22 (1.15–1.28)
Localized	5,157 (48.38)	15,585 (51.09)	0.95 (0.81–1.11)
Regional	2,385 (22.38)	6,334 (20.76)	1.08 (0.97–1.19)
Remote	1,864 (17.49)	5,567 (18.25)	0.96 (0.88–1.05)
Unknown or not specified	228 (2.14)	608 (1.99)	1.07 (0.48–2.38)
Neighborhood SES (Quintiles)			
Highest SES	3,552 (33.32)	8,703 (28.53)	1.17 (1.03–1.33)
Higher-middle SES	2,547 (23.9)	7,719 (25.30)	0.94 (0.84–1.06)
Middle SES	2,092 (19.63)	6,471 (21.21)	0.93 (0.83–1.03)
Lower-middle SES	1,513 (14.19)	4,689 (15.37)	0.92 (0.86–0.99)
Lowest SES	955 (8.96)	2,924 (9.58)	0.93 (0.91–0.96)
Payer (Insurance Type)			
Private Insurance	4,930 (46.25)	19,078 (62.54)	0.74 (0.63–0.87)

Patient Characteristic	EHR Population (N=10,659)	Non-EHR Population (N=30,506)	EHR: Non-EHR
	N (%)	N (%)	Proportion Ratio[a](95% CI [b])
Medicare	4,149 (38.92)	7,682 (25.18)	1.55 (1.37–1.74)
Any public, Medicaid, military	683 (6.41)	2,746 (9.00)	0.71 (0.69–0.73)
Not insured	73 (0.68)	216 (0.71)	0.97 (0.07–13.01)
Unknown	824 (7.73)	784 (2.57)	3.01 (1.70–5.31)
Marital Status			
Married/Reigstered	6,273 (58.85)	16,626 (54.50)	1.08 (0.92–1.27)
Single/Divorced/Widowed	3,895 (36.54)	10,165 (33.32)	1.10 (0.96–1.26)
Unknown	491 (4.61)	3,715 (12.18)	0.38 0.37–0.39)

[a] Proportion Ratio calculated as $\hat{\theta} = \frac{x_1/n_1}{x_2/n_2}$ where x_1 and x_2 are the successes in the two groups out of totals n_1 and n_2 .

[b] 95% confidence interval (CI) calculated as $\hat{\theta} exp\left[\pm 1.96\sqrt{1/x_1 - 1/n_1 + 1/x_2 - 1/n_2}\right]$

Table 3.

Use of Model-based Standardization to Reweight the EHR Population to the Covariate Distribution of the Catchment Region for the Outcome Late Stage at Diagnosis

Primary Exposure	All Catchment Region Patients (N=34,316)	EHR Population Only (N=9,158)	
	Odds Ratio (95% CI)	Unweighted Odds Ratio (95% CI)	Reweighted[e] Odds Ratio (95% CI[f])
Model 1: nSES Quintiles[a]			
Highest	REF	REF	REF
Upper Middle	1.234 (1.160–1.312)	1.275 (1.134–1.435)	1.226 (1.097–1.391)
Middle	1.431 (1.341–1.526)	1.519 (1.342–1.719)	1.421 (1.249–1.604)
Lower Middle	1.674 (1.550–1.769)	1.789 (1.560–2.051)	1.680 (1.473–1.942)
Lowest	1.958 (1.801–2.129)	2.022 (1.719–2.378)	1.990 (1.701–2.360)
Model 2: Marital Status[b]			
Married or Registered	REF	REF	REF
Single/Divorced/Widowed	1.394 (1.329–1.462)	1.373 (1.252–1.506)	1.380 (1.265–1.523)
Model 3: Primary Insurance[c]			
Private or Medicare	REF	REF	REF
Any public, Medicaid, military	1.901 (1.757–2.055)	1.844 (1.553–2.190)	1.876 (1.579–2.24)
Model 4: Patient Race/Ethnicity[d]			
Non-Hispanic White	REF	REF	REF
Non-Hispanic Black	1.135 (1.043–1.234)	1.380 (1.166–1.633)	1.130 (0.953–1.346)
Hispanic	1.184 (1.104–1.269)	1.263 (1.091–1.462)	1.189 (1.025–1.379)
Asian American/Pacific Islander	1.167 (1.100–1.238)	1.071 (0.948–1.210)	1.164 (1.028–1.307)

[a] Model adjusted for age, sex, and race/ethnicity.

[b] Model adjusted for age, sex, race/ethnicity, and nSES.

[c] Model adjusted for age, sex, race/ethnicity, nSES, and marital status.

[d] Model adjusted for age, sex.

[e] Reweighted by the stabilized weight, $sw = P(S = 1|x)/P(S = 1|y, x, \mathbf{z}, c)$, with both probabilities modeled using logistic regression, and including all two-way product terms for the denominator model.

[f] Bootstrapped confidence intervals (CI) based on 500 resamples of the EHR population (with replacement).