**1. Specific Aims**

In cancers, delays in diagnosis reduce chances of successful treatment by 60 percent [[[1]](#endnote-1)-[[2]](#endnote-2),[[3]](#endnote-3),[[4]](#endnote-4),[[5]](#endnote-5)], and increase cost of treatment by 7% to 18% [[[6]](#endnote-6),[[7]](#endnote-7)]. In breast, lung, cervical, colorectal, and prostate cancers, early detection saves lives and timely screening is important [[[8]](#endnote-8)]. The long-term goal of this project is to reduce delays in diagnosis of cancers through more effective risk-based screening.

Currently, patients are screened for cancer based on their age and gender, except for lung cancer which also relies on tobacco risk. [[9]](#endnote-9),[[10]](#endnote-10),[[11]](#endnote-11),Many have encouraged use of broader set of risk-factors in eligibility for screening [[[12]](#endnote-12)-[[13]](#endnote-13),[[14]](#endnote-14),[[15]](#endnote-15),[[16]](#endnote-16)], especially in lung [[[17]](#endnote-17),[[18]](#endnote-18)], breast [[[19]](#endnote-19),[[20]](#endnote-20)] cancers. One way to do so is through predictive models. These models include a variety of risk factors, such as: (1) lifestyle [[[21]](#endnote-21)], (2) precursor to cancer [[[22]](#endnote-22)], (3) social determinants of cancer [[[23]](#endnote-23)-[[24]](#endnote-24),[[25]](#endnote-25),[[26]](#endnote-26),[[27]](#endnote-27)], (4) chronic illnesses [[[28]](#endnote-28),[[29]](#endnote-29)], (5) viral illnesses [[[30]](#endnote-30)], and (6) other events [[[31]](#endnote-31)]. Some of these risk factors are causes of cancers; others are only statistically associated with cancers. Predictive models may rely on thousands of medical history variables, mixing causes with associated risk factors. The mix and number of risk factors included in the predictive models make this approach from current practice of narrowly defined eligibility requirements for screening.

The US Preventive Services Task Force (USPSTF) recommends the timing of cancer screening. Providers and insurance companies often adopt the recommendations, highlighting the importance of this organization in changing clinical practices [[[32]](#endnote-32)-,[[33]](#endnote-33),[[34]](#endnote-34)]. USPSTF recommends eligibility for cancer screening, primarily based on the inclusion/exclusion criteria of random clinical trials for screening interventions. In these trials, the inclusion criteria are narrowly defined to minimize the needed sample size. Not surprisingly, the related USPSTF recommendations are also narrowly defined. In contrast, predictive models are based on observational data from electronic health records (EHRs). These models are high-dimensional, pragmatic, encompass all races, and can rely on thousands of medical history variables. USPSTF has not included predictive models in any of its recommendations. The consequences of USPSTF ignoring this literature are not well understood, as there have been no comparisons of current USPSTF recommendations and predictive model screening. This study exposes how many lives are lost because of USPSTF line of reasoning. In particular:

1. **The first aim of this study is to develop and compare the accuracy of EHR-based predictive models for breast, lung, cervical, colorectal, and prostate cancers to USPSTF’s recommendations.** We plan to use EHR data from 519,000 participants in the NIH-funded “All of Us” project [[[35]](#endnote-35)]. We will report the cross-validated accuracy of these predictive models, and compare their accuracy to the USPSTF’s age-and-gender eligibility recommendations. We hypothesize that if screen eligibility requirements were based on predictive models, both missed cancers and false alarms will be reduced. We will test this hypothesis using area under the receiver operating characteristic curve.
2. **The second aim of this study is to measure extent of delay associated with the two approaches for setting eligibility for cancer screening.** To accomplish this, we trace the day of eligibility for both approaches and use right-censored survival analysis to report the extent of the delay. Predictive models may make patients eligible months before USPSTF does so.

**Impact:** This project is a demonstration of the potential of big data and predictive medicine to improve cancer screening [[[36]](#endnote-36)]. Predictive models, while accurate, are seldom used in clinical care [[[37]](#endnote-37)]. The use of a predictive model requires shifting from the current narrowly defined, causal, easily understood, USPSTF’s screening recommendations to complex, risk-based, comprehensive, not-necessarily causal, machine/EHR-dependent, predictive models. This is not an easy shift in clinical practices and is unlikely to occur without the study of the implications of the shift. By laying out the consequences of this shift in terms of lives saved, this study helps patients, clinicians, insurance companies, and of course USPSTF make a well-informed choice.

**Significance**

This proposal fulfills a vision for the future of healthcare, in which predictive medicine, and big data, are central to clinical practices. Many have anticipated these changes and called it inevitable [[[38]](#endnote-38)] but to date predictive medicine has not been put to use in cancer screening [[[39]](#endnote-39)]. This project delivers on promises of predictive medicine and introduces a new way for risk-based cancer screening.

This study is significant because it quantifies the extent of diagnostic delay in the current cancer screening programs, heavily influenced by USPSTF’s recommendations [[[40]](#endnote-40)-,[[41]](#endnote-41),[[42]](#endnote-42)]. Diagnostic delays have devastating consequences [[[43]](#endnote-43)]. Delays in diagnosis of cancer reduce chances of successful treatment by 60 percent [[[44]](#endnote-44)-[[45]](#endnote-45),[[46]](#endnote-46),[[47]](#endnote-47),[[48]](#endnote-48)], and increase cost of treatment by 7% to 18% [[[49]](#endnote-49),[[50]](#endnote-50)]. Once the impact of diagnostic delay is better understood, USPSTF may modify its procedures and rely more on predictive models. The significance of this project does not entirely depend on USPSTF’s adoption of these methods. Once the alternative method of establishing eligibility is freely available on the web, then patients and clinicians can start using it through EHRs or through online calculators. This project is significant because it provides EHR vendors, clinicians, and patients with a convenient, comprehensive, non-invasive, universal, and free method for assessing eligibility for cancer screening.

For years, providers have called for more risk-based screening [[[51]](#endnote-51)]. Current approaches to risk have been limited and selective. This project expands what is considered a legitimate cancer risk factor. This project includes any aspect of the patient’s medical history as a risk factor (see the Innovation Section for a literature review). This project is significant because for the first time, it expands the list of risk factors for cancer, with no consideration that these factors must cause cancer; mere statistical association is sufficient to elevate risk. Of course, modification of non-causal risk factors does not reduce risk of cancer but these modifications are not the goal of our planned risk models.

This study is also significant because it shows how social determinants of cancers can increase timely screening and reduce diagnostic delays. Many argue that eligibility for screening should be based on social determinants [[[52]](#endnote-52)] because the burden of diagnostic delays falls disproportionately on communities of color and underserved populations [[[53]](#endnote-53),[[54]](#endnote-54)]. EHR-based predictive models use race, ethnicity, and social factors (Z-codes) within electronic health records to construct more accurate risk factors for cancer. USPSTF’s recommendations do not take into account these factors. This study is significant because it makes screening recommendations that improve health equity.

**Strategy**

**Innovation**

We are responding to the request for proposals: Secondary Analysis and Integration of Existing Data to Elucidate Cancer Risk and Related Outcomes. This study leverages and performs innovative analyses of existing data. We propose high-dimensional predictive models for determining eligibility for (1) breast, (2) lung, (3) cervical, (4) colorectal, and (5) prostate cancer screens. USPSTF reviews indicate that only in these cancers early detection can lead to more effective treatment [[[55]](#endnote-55)]. There is also good reason to expect that our proposed EHR-based high-dimensional predictive models for cancers can be more accurate than USPSTF’s recommendations for the same cancers. The USPSTF recommendations are primarily focused on age and gender and ignore a variety of known risk factors (see **Table 1**).

**Table 1: Recommendations of USPSTF & Missed Risk Factors**

|  |
| --- |
| **Cervical Cancer:** For Pap smear, women aged 21 to 65 years; for CHP testing women 30 to 65 years. **Missed Known Risk Factors:** (a) sexually transmitted diseases, (b) diseases and medications indicating weakened immune system, (c) exposure to smoking or smoking, (d) use of oral contraceptives, (e) multiple child births, and (f) obesity [[[56]](#endnote-56),[[57]](#endnote-57)]. |
| **Breast Cancer:** Biennial mammography screens for women aged 50 to 74 years. Additional screening for women aged 40-49 years based on clinical evaluation of risks. **Missed Known Risk Factors:** For women age 50 to 74, (a) obesity, (b) oral contraceptives use, (c) early menarche, (d) late menopause, (e) late age at first pregnancy, (f) low pregnancy parity, (g) hormone replacement therapy, (h) smoking or exposure to smoking, (i) alcoholism, and (j) certain lifestyle diseases [[[58]](#endnote-58),[[59]](#endnote-59)]. |
| **Lung Cancer:** Adults aged 50 to 80 years who have a 20 pack-year smoking history and currently smoke or have quit within the past 15 years. **Missed Known Risk Factors:** (a) exposure to radon, (b) exposure to environmental pollution, (c) certain occupational exposures, (d) gender, (e) race, and (f) a variety of pre-existing lung diseases [[[60]](#endnote-60),[[61]](#endnote-61)].  |
| **Colorectal Cancer:** Adults aged 45 to 75 years. **Missed Known Risk Factors:** (a) obesity, (b) low physical activity, (c) smoking, (d) high salt diet, and (e) red meat consumption [[[62]](#endnote-62)]. |
| **Prostate Cancer:** Clinicians should evaluate the need for PSA tests. **Missed Risk Factors**: Depends on clinicians. |

In cancer management, one form of predictive models, often referred to as web calculators, is in use [[[63]](#endnote-63)-[[64]](#endnote-64)[[65]](#endnote-65)[[66]](#endnote-66)[[67]](#endnote-67)] and no longer innovative. In recent years, with widespread availability of EHRs, investigators have proposed models that predict risk of lung [[[68]](#endnote-68)-,[[69]](#endnote-69),[[70]](#endnote-70)], breast [[[71]](#endnote-71)-,[[72]](#endnote-72),[[73]](#endnote-73)], cervical [[[74]](#endnote-74)], and colorectal [[[75]](#endnote-75)-,[[76]](#endnote-76),[[77]](#endnote-77)] cancer. A recent review shows that these predictive models use a limited number of variables, less than 100 variables [[[78]](#endnote-78),[[79]](#endnote-79)] and can be called selective predictive models. In contrast, we propose predictive models that include thousands of comorbidities, procedures, and medications, such as:

1. *Obvious Predictors*. Since cancers can metastasize to other organs, incorporating a history of previous cancer is an obvious predictor. None of the existing selective predictive models do so.
2. *Rare Diseases*. We include rare diseases as risk factors for cancers. For example, Li-Fraumeni syndrome increases risk of cancer by 50 folds [[[80]](#endnote-80)]. Rare diseases are not part of the selective models, even though a quarter of all diseases are rare. “Although rare diseases are individually rare by definition, collectively they are common” [[[81]](#endnote-81)]. The cumulative effect could be large [[[82]](#endnote-82)].
3. *Precursors to Cancers*. Contrary to selective models, we include cancer precursors, e.g., COPD for lung cancer [[[83]](#endnote-83)], in our models.
4. *Viruses.* A variety of viruses (e.g., Epstein-Barr, Hepatitis B, Hepatitis C, HIV) can cause cancers [[[84]](#endnote-84),[[85]](#endnote-85)]. Selective models do not take into account risks caused by viruses. We do.
5. *Chronic Diseases*. We use chronic diseases (e.g., cardiovascular disease, diabetes, chronic kidney disease, COPD) to predict cancer risks, but selective approaches do not. Chronic diseases are more important markers for cancers than smoking, diet, alcohol and obesity combined [[[86]](#endnote-86)].
6. *Other Diseases*. We include other predictors of cancers (e.g., vitamin k-deficiency [[[87]](#endnote-87)] and tuberculosis [[[88]](#endnote-88)] for predicting lung cancer), even though the link is not causal or well understood.

**Approach**

**Preliminary Analyses**

We examined if lung cancer can be predicted from patient’s medical history within EHRs. For this analysis, we used data available to us through the Veterans Affairs Informatics and Computing Infrastructure. Nearly complete longitudinal medical records were available for 8.7 million veterans. The dependent variable was a diagnosis of lung and bronchus cancer, as identified using the Agency for Healthcare Research and Quality (AHRQ)'s Clinical Classification System. The independent variables were thousands of diagnostic codes. To avoid modeling noise, data were divided randomly into a training dataset (90% of the cases) and a validation dataset (10% of the cases). Composite variables were constructed using the likelihood ratio associated with each diagnosis, $LR\_{Dx}$:

$$LR\_{Dx}=\frac{p\left(Cancer 6 months after date of diagnosis\right)}{p\left(No cancer 6 months after date of diagnosis\right)}$$

Certain diagnoses, such as "Deficiency of vitamin K," "Lack of housing," "Sialadenitis," "Infection of tracheostomy," and "BMI 60.0-69.9," were associated with increased likelihood of lung cancer.

Certain diagnoses, such as "Deficiency of vitamin K," "Lack of housing," "Sialadenitis," "Infection of tracheostomy," and "BMI 60.0-69.9," were associated with increased likelihood of lung cancer. We grouped the diagnoses into 11 body systems and 1 system for social determinants of illness. In each system, we use the worst diagnosis The systems we examined were: (1) allergies, (2) autoimmune disorders, (3) cardiac conditions, (4) dermatologic issues, (5) endocrine disorders, (6) gastrointestinal

|  |
| --- |
| **Table 2: Examples of Body System Predictors** |
| **Likelihood Ratio** | **Infectious & Parasitic Diseases Body System** |
| **1 to 1.99** | Salmonella infection, unspecified |
| **2 to 2.99** | Chronic Infection of amputation stump |
| **3 to 3.99** | Infection resistant to penicillin |
| **4 to 4.99** | Hemophilus influenzae H. infection |
| **5 to 5.99** | Infection of gastrostomy |
| **6 to 6.99** | Intestinal infection due to clostridium difficile |
| **7 to 7.99** | Infection resistant to cephalosporins |
| **8 to 8.99** | Friedländer's bacillus infection |
| **9 to 9.99** | Infection of tracheostomy |
| **10+** | Viral hepatitis C without hepatic coma |

problems, (7) hematologic disorders, (8) immunodeficiency, (9) infections, (10) neurologic conditions, (11) social determinants of illness, and (12) other oncological diseases besides lung cancer. **Table 2** shows selected disease markers for 1 out of the 12 systems. For example, “Salmonella” was a relatively benign disease within infectious diseases, it did not increase the odds of lung cancer by much. In contrast, “Viral hepatitis C” increased the odds of a later lung cancer by more than 10 folds. In the “Infection and Parasitic Diseases” body system, the worst disease was used to predict risk of cancer and all other diseases were ignored.

**Figure 2** shows the results. In cross-validation, the model had an AROC of 0.87. This pilot confirms that the method of constructing composite risk factors is reasonable and predictive of lung cancer risks. Others have also constructed similar models [[[89]](#endnote-89)]. The final method we are planning to use is a Probability Network model, only a portion of this network is modeled in this preliminary study. This pilot models the main model, which links the risk factors to lung cancer. The network also includes other models used to predict missing values for risk factors of lung cancer.

|  |
| --- |
| **Figure 2: Predicting Lung Cancer** A picture containing text, screenshot, line, plot  Description automatically generated |

**Study Design:** This is a retrospective, observational, case-control study, where cases are patients with cancer (identified through SNOMED codes) and controls are non-cancer patients. Separate models are built for each cancer: (1) breast, (2) lung, (3) cervical, (4) colorectal, and (5) prostate cancers

**Study Population and Sample Size**: We will rely on data available through All of Us Research Hub, coordinated by the National Institute of Health [[[90]](#endnote-90)]. As of January 2024, the demographic breakdown of participants was: 64.3% female, 18.7% African American, 21.1% other races, and 60.1% White. 16.9% were Hispanic. Participants’ age ranged from 18 to 100 years. As of January 2024, using only the EHR data, there were 50,720 (out of 519,000 unique participants) who had cancer [[[91]](#endnote-91),[[92]](#endnote-92)]. All participants are included in the analysis independent of their health, as long as they have at least 1 year of history. Those with less than 1 year of history will be analyzed in the second year of this study.

**Methods of Analysis of Aim 1:** For Aim 1, the primary outcome of interest is the presence/absence of the cancers. The independent variables (risk factors) are diseases in the medical history of the patients. These diseases are identified using 25,000 SNOMED codes. The study will also analyze social determinants of cancer, measured also through SNOMED codes, stemming from Z-codes [[[93]](#endnote-93),[[94]](#endnote-94)]. According to published data, these codes are available between 2% to 12% of cases [[[95]](#endnote-95),[[96]](#endnote-96),[[97]](#endnote-97)]. SNOMED codes are used to construct the worst disease in the body system. As the preliminary study has demonstrated, these constructed composite features are highly predictive of cancer. We plan to use Probability Networks, also later used in Causal Networks [[[98]](#endnote-98),[[99]](#endnote-99)] to model the data. We are not making any causal claims. We prefer this modeling approach because:

1. *Transparent Model.* In contrast to deep learning and Neural Networks, Probability Networks do not rely on a black box of interactions among the variables [[[100]](#endnote-100)].
2. *Temporal Constraints*. Probability Networks are constrained by the temporal order of predictors of cancer. This information is widely available in EHRs (as every event is timestamped) and can increase accuracy [[[101]](#endnote-101)]. None of the existing data mining techniques use the timing of variables.
3. *Redundancy*. Probability Network models are more robust because they include both direct and indirect predictors of cancers, predictors that are intercorrelated.
4. *Removal of Confounding*. Network models remove confounding in the data through inverse propensity weights [[[102]](#endnote-102)] and stratification [[[103]](#endnote-103),[[104]](#endnote-104)], procedures not easily available in other methods.
5. *Missing Values*. Networks are a collection of statistical models. The main model relates predictors, or risk factors, to cancer. Networks also include auxiliary models that impute missing risk factors in the main model. In this sense, network models easily adjust to incomplete medical records.
6. *Interactions*. Like other data mining techniques, network models allow interactions among the variables and can produce complex but accurate models [[[105]](#endnote-105),[[106]](#endnote-106),[[107]](#endnote-107)].

We plan to construct Networks through repeated, temporally-constrained, chains of Least Absolute Shrinkage and Selection Operator (LASSO) regressions. Alemi, the PI in this project, has also shown how chain of LASSO regressions can identify both the structure and the parameters of the Network [[[108]](#endnote-108)]. It is called a “chained regression” because the statistically significant variables in one regression serve as response variables for the next set of regressions. First, the cancer is LASSO regressed on temporally constrained independent variables, identifying direct predictors of the cancer. Next, direct predictors are regressed on independent variables that occur prior to it, thereby identifying indirect predictors of the cancer. As the chain of regressions continues, the response variable occurs earlier in time, and the risk factors that precede it become fewer. The chain stops when no risk factors remain.

To reduce the possibility of modeling noise, LASSO regressions will be based on 10-fold cross- validation [[[109]](#endnote-109)]. To ensure that findings are robust, the data will be sampled 40 times and variables that show in 90% of LASSO regressions will be used as risk factors [[[110]](#endnote-110)]. Once a parsimonious set of risk factors is identified, these variables are entered into a parametric logistic regression to get the final estimates of the coefficients [[[111]](#endnote-111)].

Separate models are built for each of the five cancers. Multiple test adjustments will be applied. Model parameters are derived from the training data set and cross-validated on a test set. The 10-fold cross-validated accuracy of predictions will be tested using Area under the Receiver Operating Characteristic curves (as also done in our preliminary study). In addition, we will also look at calibration using Brier scores [[[112]](#endnote-112)], reclassification tables [[[113]](#endnote-113)], and net reclassification improvement [[[114]](#endnote-114)].

In addition to individual models for each cancer, the following procedures will be used to predict all 5 cancers within the same model. If $y\_{ijt}$ indicates the presence of cancer for participant i at time t clustered in cancer type j, then risk factors for the cancers will be examined using the following model:

$$logit(y\_{ijt})=β\_{0}+\sum\_{r}^{}β\_{r}x\_{jtr}+\sum\_{r}^{}φ\_{t}x\_{jtr}+ν\_{0ij}+ε\_{ijt},$$

Where $β\_{0}$ is the intercept, $β\_{r}$ is the risk factor effect, $x\_{jtr}$ is the indicator for whether or not cancer type $j$ is affected by the risk factor $r$, $φ\_{t}$ is the estimate of the additional effect of the cancer diagnosis at time $t$. This model accounts for both a step change in risk of cancer and a change in the risk over calendar time, which is specified by including the $φ\_{t}x\_{jtr}$ interaction between the calendar time and risk factor in the model. $ν\_{0ij}$ is a random effect term, which can be expressed as $ν\_{0j}+h\_{0i}$, where $ν\_{0j}$ is a cluster specific random intercept and $h\_{0i}$ is a random term for the $i$th patient. Finally, $ε\_{ijt}$ is the random error term.

The project has access to sufficient data to develop the proposed models. Assuming a medium effect size of 0.25 and a significance level of 0.05, a simulation-based power analysis was performed in RStudio, and a power of 80% or more was estimated for varying levels of correlation among observations over time (low to medium). Using the approaches of Muller and Glueck [[[115]](#endnote-115),[[116]](#endnote-116)], a covariance matrix was added to account for repeated measures within the final model. The available data of 519,000 cases provide sufficient power of at least 0.80 for detecting the five cancer diagnoses. LASSO models will be well-powered, given that LASSO is known to reduce dimensionality and requires a smaller sample compared to linear models [[[117]](#endnote-117)].

Sufficient data are also available to examine models within racial and gender subgroups. Separate models will be developed for women, African Americans and Hispanics. The majority (78%) of participants in All of Us project have traditionally been under-represented in NIH funded studies, 49% are not White; and compared to national rates, African Americans are over-sampled [[[118]](#endnote-118)].

**Methods of Analysis of Aim 2:** For Aim 2, the primary outcome of interest is days of delay in eligibility for cancer screening, measured as: $T=min⁡(C,D)-min⁡(P,D)$, where C is the day patient is eligible for screening based on USPSTF recommendations and P is the day of eligibility based on risk-based predictive models, and D is the day data are censored because of death or end of the follow-up period. The delay in eligibility is not the same as a delay in screening, which depends on a variety of factors such as screening technology, raised awareness of cancer symptoms, patient characteristics, provider practice patterns, insurance and cost. In this proposal, we focus on delay in eligibility, assuming that other factors have an impact on actual screening but that such an effect will be the same if eligibility is set by predictive models versus USPSTF’s recommendations. In Aim 2, the delay in eligibility is recognized as the contrast of interest. The way we define this contrast is similar to how others have done so for comparing two screening tests [[[119]](#endnote-119)-,[[120]](#endnote-120),[[121]](#endnote-121),[[122]](#endnote-122)]. The reliance on eligibility contrast is attractive because it isolates the impact of USPSTF’s recommendations separate from other contributors to delay, it is easy to obtain, and it provides a lead time bias that can also affect actual screening times.

We plan to use the Kaplan–Meier estimator, to estimate the days of diagnostic delay (as a survival function) from the data [[[123]](#endnote-123)]. The Kaplan–Meier estimator is a valuable tool for estimating the duration of delay [[[124]](#endnote-124)]. This is particularly relevant, where the delay can be observed through an overlap in the Kaplan–Meier curves for the two groups [[[125]](#endnote-125)]. Test statistics have been proposed for comparing Kaplan–Meier curves to delayed effects [[[126]](#endnote-126)]. This method of examining delays is practical because it highlights the adverse effects of delays on tumor control [[[127]](#endnote-127)]. The estimator of the survival function, $\hat{S}\left(t\right)$, i.e., the probability that delay is longer than t, is given by:

$$\hat{S}\left(t\right)=\prod\_{i:t\_{i}\leq t}^{}\left(1-\frac{d\_{i}}{n\_{i}}\right)$$

�Where $t\_{i}$ is the �^(�)=∏�: ��≤�(1−����),time when at least one person has a delay of more than t days, di is the number of individuals in whom the delay was longer than t days��, and $n\_{i}$�� is the number of individuals who have not yet had a delay of t days.

��The Kaplan–Meier curve can take into account right censored data [[[128]](#endnote-128)], which occurs if a patient dies, or is alive without either C or P occurring before last entry into the EHR.

Most USPSTF’s recommendations are based on age and gender, which are readily available in EHRs , and therefore, it is relatively easy to establish the day the patient was eligible for screening. Some USPSTF recommendations use constructs that are not coded within EHRs. For example, USPSTF recommendations for lung cancer screening requires “20 pack-year smoking history” [[[129]](#endnote-129)]. This information is not available in EHR codes [[[130]](#endnote-130)] and must be imputed from available data. A number of investigators have shown this could be done from notes [[[131]](#endnote-131)] or from structured EHR data [[[132]](#endnote-132)]. In All of Us data, 398,200 participants have completed a lifestyle survey that includes detailed nicotine intake data. These data will be used to impute missing nicotine exposure in the EHR data.

In addition, some USPSTF eligibility recommendations rely on the clinician’s independent risk assessment. For example, eligibility for mammograms at ages 35-49 relies on the providers assessment of the risk of breast cancer. USPSTF recommends the assessment but does not specify how it should be done. In age groups where USPSTF recommends clinical assessment, we plan to rely on the date of screening interventions to set an earlier eligibility date.

Predictive models produce a continuous scale. To understand at what risk the patient should be screened for cancer one has to have a cutoff value, above which patients are considered to be at elevated risk. Two approaches have been reported in the literature: (1) patients whose risk exceeds the prevalence in the dataset may be classified as high-risk, (2) patients whose risk exceeds the optimal cut point, established through finding a threshold that minimizes the sum of the number of false positive and false negative cancers [[[133]](#endnote-133)]. We prefer the latter but will conduct sensitivity analysis to see if study findings depend on the optimal cutoff point.

For power estimation of Aim 2, a range of hazard ratios have been considered using different cancer screening studies. In general, small, medium, and large hazard ratios would be approximately 1.3, 1.9, and 2.8, respectively. In varying types of cancer, screening studies report ratios between 0.7 and 1.6 [[[134]](#endnote-134)-[[135]](#endnote-135),[[136]](#endnote-136)]. Using this hazard ratio, a sample size of 1,500 will give us at least 80% power. Considering the most conservative scenario with a 1.1 hazard ratio (which is a lot more conservative than any reported hazard ratios in the cancer screening studies), with a sample size between 8,600 and 11,400, we will be powered at 0.8 and 0.9, respectively. Considering that our sample size is higher than this (in January 2024, it was 50,720), we will be powered at over 90% to complete Aim 2 of this study.

**Team’s Roles & Expertise.** Farrokh Alemi, Ph.D. will serve as the Principal Investigator, responsible for all aspects of the project. He is an experienced data scientist, working with massive data. He is also a leading expert in the construction of Probability Networks. Kyung Lee, Ph.D. will serve as a data scientist (Co-I). Dr. Lee will primarily be involved in developing predictive models using Network-based mediation analysis. Niloofar Ramezani Ph.D., a statistician (Co-I), will compare the performance of predictive models to the recommendations of USPSTF. Hoda Bidkhori, Ph.D., will also serve as a data scientist (Co-I) and focus on alternatives to Probability Network models. Mark Schwartz, MD, is an experienced primary care physician and researcher (Co-I). He will help interpret findings and their suitability for use in primary care practices. Xie Wang, MD, PhD, is an experienced cancer researcher focused on cancer screening. She will provide insights into cancer screening process. Dr. Alemi has 7 publications with Dr. Schwartz, 3 publications with Dr. Lee, and 1 conference paper with Dr. Ramezani. Dr. Bidkhori and Dr. Wang are new to the team. Most members of the team have successful history of working together.

**Risks, Limitations, and Mitigation.** This study assumes that individuals will be screened on the day they become eligible through predictive modeling. Of course, such an assumption is false as individuals may not follow through and get screened. Some patients may self-screen earlier; others may delay screening. Provider practice patterns may also affect screening days. Medicare’s or insurance companies’ reimbursement for screening may also affect the actual day of screening. We acknowledge that a myriad of patient, provider, or organizational factors affect screening. We assume that these factors will continue to affect screening independent of how eligibility was established. The goal of this project is to highlight how the recommendations themselves affect eligibility for screening and not how other factors lead to delay in screening, once one is eligible. Of course, a delay in eligibility is not the same as a delay in cancer diagnosis. At the same time, data show that eligibility recommendations affect the actual detection of cancers [[[137]](#endnote-137)-[[138]](#endnote-138),[[139]](#endnote-139),[[140]](#endnote-140),[[141]](#endnote-141)]. The focus on eligibility has the advantage of highlighting the direct role of USPSTF in the detection of cancers, separate from patient, provider, and organizational causes of delay in screening.

The study plans to use Probability Networks for modeling purposes. If this method of analysis is not accurate enough, we will adjust our plans, and examine Deep Learning (Neural Networks) or Support Vector Machines to improve the accuracy of our modeling efforts.

**References**

1. Shapiro S, Strax P, Venet L. Periodic breast cancer screening in reducing mortality from breast cancer. Jama. 1971; 215:1777-85. [↑](#endnote-ref-1)
2. Chad‐Friedman E, Coleman S, et al. Psychological distress associated with cancer screening: a systematic review. Cancer. 2017; 123: 3882-94. [↑](#endnote-ref-2)
3. Sarma EA, Silver MI, Kobrin SC, Marcus PM, Ferrer RA. Cancer screening: health impact, prevalence, correlates, and interventions. Psychology & health. 2019; 34:1036-72. [↑](#endnote-ref-3)
4. Lin JS, Piper MA, et al. Screening for colorectal cancer: updated evidence report and systematic review for the US Preventive Services Task Force. JAMA. 2016; 315:2576-94. [↑](#endnote-ref-4)
5. Zarcos-Pedrinaci I, Téllez T, et al. Factors associated with prolonged patient-attributable delay in the diagnosis of colorectal cancer. Cancer Research and Treatment. 2018; 50:1270-80 [↑](#endnote-ref-5)
6. Maringe C, Walters S, et al. Stage at diagnosis and colorectal cancer survival in six high-income countries: a population-based study of patients diagnosed during 2000–2007. Acta oncologica. 2013; 52:919-32. [↑](#endnote-ref-6)
7. Knaul FM, Farmer PE, et al. Alleviating the access abyss in palliative care and pain relief—an imperative of universal health coverage: the Lancet Commission report. The Lancet. 2018; 391:1391-454. [↑](#endnote-ref-7)
8. Andersen BL, Cacioppo JT, Roberts DC. Delay in seeking a cancer diagnosis: delay stages and psychophysiological comparison processes. British journal of social psychology. 1995; 34:33-52. [↑](#endnote-ref-8)
9. Plambeck BD, Wang LL, et al. Effects of the 2012 and 2018 US preventive services task force prostate cancer screening guidelines on pathologic outcomes after prostatectomy. Prostate. 2022 Feb;82(2):216-220. [↑](#endnote-ref-9)
10. Maki KG, Talluri R, et al. Impact of U.S. Preventive Services Task Force lung cancer screening update on drivers of disparities in screening eligibility. Cancer Med. 2023 Feb; 12(4): 4647-4654. [↑](#endnote-ref-10)
11. Fleshner K, Carlsson SV, Roobol MJ. The effect of the USPSTF PSA screening recommendation on prostate cancer incidence patterns in the USA. Nat Rev Urol. 2017 Jan;14(1):26-37. [↑](#endnote-ref-11)
12. Ten Haaf K, Jeon J, et al. Risk prediction models for selection of lung cancer screening candidates: A retrospective validation study. PLoS Med. 2017 Apr 4;14(4):e1002277. [↑](#endnote-ref-12)
13. Rivera MP, Katki HA, et al. Addressing disparities in lung cancer screening eligibility and healthcare access. An Official American Thoracic Society Statement. Am J Respir Crit Care Med. 2020 Oct 1; 202(7): e95-e112. [↑](#endnote-ref-13)
14. Gail MH, Brinton LA, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst. 1989 Dec 20;81(24):1879-86. [↑](#endnote-ref-14)
15. Benichou J. A computer program for estimating individualized probabilities of breast cancer. Comput Biomed Res. 1993 Aug;26(4):373-82. [↑](#endnote-ref-15)
16. Beyer K, Leenen R, et al. Health policy for prostate cancer early detection in the European Union and the impact of opportunistic screening: PRAISE-U Consortium. J Pers Med. 2024 Jan 11;14(1):84. [↑](#endnote-ref-16)
17. Aredo JV, Choi E, et al. Racial and Ethnic Disparities in Lung Cancer Screening by the 2021 USPSTF Guidelines Versus Risk-Based Criteria: The Multiethnic Cohort Study. JNCI Cancer Spectr. 2022 May 2;6(3):pkac033. [↑](#endnote-ref-17)
18. Toumazis I, Bastani M, et al. Risk-Based lung cancer screening: A systematic review. Lung Cancer. 2020 Sep;147:154-186. [↑](#endnote-ref-18)
19. Harkness EF, Astley SM, Evans DG. Risk-based breast cancer screening strategies in women. Best Pract Res Clin Obstet Gynaecol. 2020 May;65:3-17. [↑](#endnote-ref-19)
20. Lee CI, Chen LE, Elmore JG. Risk-based Breast Cancer Screening: Implications of Breast Density. Med Clin North Am. 2017 Jul;101(4):725-741. [↑](#endnote-ref-20)
21. Lanting CI, de Vroome EM, et al. [Contribution of lifestyle factors to cancer: secondary analysis of Dutch data over 2010 and a projection for 2020]. Ned Tijdschr Geneeskd. 2014;159:A8085. [↑](#endnote-ref-21)
22. Wacholder S. Precursors in cancer epidemiology: aligning definition and function. Cancer Epidemiol Biomarkers Prev. 2013 Apr;22(4):521-7. doi: 10.1158/1055-9965.EPI-13-0084. [↑](#endnote-ref-22)
23. Fallahian F, Nelson H, Pories S. The role of social determinants of health on cancer screening. CLINICAL CONGRESS 2022; 2022. [↑](#endnote-ref-23)
24. Choy AM, Lebwohl B, et al. Impact of social determinants of health on colorectal cancer screening and surveillance in the COVID reopening phase. European Journal of Gastroenterology & Hepatology. 2022; 34: 739. [↑](#endnote-ref-24)
25. Mohan G, Chattopadhyay S. Cost-effectiveness of leveraging social determinants of health to improve breast, cervical, and colorectal cancer screening: a systematic review. JAMA oncology. 2020; 6:1434-44. [↑](#endnote-ref-25)
26. Zettler ME, Feinberg BA, Jeune-Smith Y, Gajra A. Impact of social determinants of health on cancer care: a survey of community oncologists. BMJ open. 2021; 11:e049259. [↑](#endnote-ref-26)
27. Lofters A, Schuler A, et al. Using self-reported data on the social determinants of health in primary care to identify cancer screening disparities. BMC Family Practice. 2017; 18:1-11. [↑](#endnote-ref-27)
28. Tu H, Wen CP, et al. Cancer risk associated with chronic diseases and disease markers: prospective cohort study. BMJ. 2018 Jan 31;360:k134. [↑](#endnote-ref-28)
29. Brenner DR, McLaughlin JR, Hung RJ. Previous lung diseases and lung cancer risk: a systematic review and meta-analysis. PLoS One. 2011 Mar 31;6(3):e17479. [↑](#endnote-ref-29)
30. Dalton-Griffin L, Kellam P. Infectious causes of cancer and their detection. J Biol. 2009;8(7):67. [↑](#endnote-ref-30)
31. Wu S, Zhu W, et al. Evaluating intrinsic and non-intrinsic cancer risk factors. Nat Commun 2018, 9, 3490. [↑](#endnote-ref-31)
32. Qian Z, Chen X, et al. Changes in prostate-specific antigen screening after the 2018 United States Preventive Services Task Force Recommendations and through the COVID-19 pandemic. Eur Urol Oncol. 2024 Feb; 7(1): 151-154. [↑](#endnote-ref-32)
33. Khairnar R, Mishra MV, Onukwugha E. Impact of United States Preventive Services Task Force Recommendations on utilization of prostate-specific antigen screening in Medicare beneficiaries. Am J Clin Oncol. 2018 Nov;41(11):1069-1075. [↑](#endnote-ref-33)
34. Ali MK, Imperatore G, et al. Impact of changes in diabetes screening guidelines on testing eligibility and potential yield among adults without diagnosed diabetes in the United States. Diabetes Res Clin Pract. 2023 Mar;197:110572. [↑](#endnote-ref-34)
35. See https://www.researchallofus.org/data-tools/data-snapshots/ accessed on 2/2/2024 [↑](#endnote-ref-35)
36. Parikh RB, Kakad M, Bates DW. Integrating predictive analytics into high-value care: the dawn of precision delivery. JAMA. 2016 Feb 16;315(7):651-2. [↑](#endnote-ref-36)
37. Yang C, Kors JA, et al. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. Journal of the American Medical Informatics Association. 2022 May 1; 29(5): 983-9. [↑](#endnote-ref-37)
38. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA. 2013 Apr 3;309(13):1351-2. doi: 10.1001/jama.2013.393. PMID: 23549579. [↑](#endnote-ref-38)
39. Yang C, Kors JA, et al. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. Journal of the American Medical Informatics Association. 2022 May 1; 29(5): 983-9. [↑](#endnote-ref-39)
40. Qian Z, Chen X, et al. Changes in prostate-specific antigen screening after the 2018 United States Preventive Services Task Force Recommendations and through the COVID-19 pandemic. Eur Urol Oncol. 2024 Feb; 7(1): 151-154. [↑](#endnote-ref-40)
41. Khairnar R, Mishra MV, Onukwugha E. Impact of United States Preventive Services Task Force recommendations on utilization of prostate-specific antigen screening in Medicare beneficiaries. Am J Clin Oncol. 2018 Nov; 41(11): 1069-1075. [↑](#endnote-ref-41)
42. Ali MK, Imperatore G, et al. Impact of changes in diabetes screening guidelines on testing eligibility and potential yield among adults without diagnosed diabetes in the United States. Diabetes Res Clin Pract. 2023 Mar; 197:110572. [↑](#endnote-ref-42)
43. Allen B, Chatfield M, Burleson J, Thorwarth WT. Improving diagnosis in health care: perspectives from the American College of Radiology. Diagnosis. 2017; 4:113-24. [↑](#endnote-ref-43)
44. Shapiro S, Strax P, Venet L. Periodic breast cancer screening in reducing mortality from breast cancer. JAMA 1971; 215:1777-85. [↑](#endnote-ref-44)
45. Chad‐Friedman E, Coleman S, et al. Psychological distress associated with cancer screening: a systematic review. Cancer. 2017; 123: 3882-94. [↑](#endnote-ref-45)
46. Sarma EA, Silver MI, Kobrin SC, Marcus PM, Ferrer RA. Cancer screening: health impact, prevalence, correlates, and interventions. Psychology & health. 2019; 34:1036-72. [↑](#endnote-ref-46)
47. Lin JS, Piper MA, et al. Screening for colorectal cancer: updated evidence report and systematic review for the US Preventive Services Task Force. JAMA. 2016; 315:2576-94. [↑](#endnote-ref-47)
48. Zarcos-Pedrinaci I, Téllez T, et al. Factors associated with prolonged patient-attributable delay in the diagnosis of colorectal cancer. Cancer Research and Treatment. 2018; 50:1270-80 [↑](#endnote-ref-48)
49. Maringe C, Walters S, et al. Stage at diagnosis and colorectal cancer survival in six high-income countries: a population-based study of patients diagnosed during 2000–2007. Acta oncologica. 2013; 52:919-32. [↑](#endnote-ref-49)
50. Knaul FM, Farmer PE, et al. Alleviating the access abyss in palliative care and pain relief—an imperative of universal health coverage: the Lancet Commission report. The Lancet. 2018; 391:1391-454. [↑](#endnote-ref-50)
51. Laza-Vásquez C, Codern-Bové N, et al. Views of health professionals on risk-based breast cancer screening and its implementation in the Spanish National Health System. PLoS One. 2022 Feb 4; 17(2): e0263788. [↑](#endnote-ref-51)
52. O'Brien MJ, Kirley KA, Ackermann RT. Reducing health disparities through prevention: role of the U.S. Preventive Services Task Force. Am J Prev Med. 2020 May; 58(5): 724-727. [↑](#endnote-ref-52)
53. Esnaola NF, Ford ME. Racial differences and disparities in cancer care and outcomes: where’s the rub? Surgical Oncology Clinics. 2012; 21:417-37. [↑](#endnote-ref-53)
54. Mishra SI, DeForge B, et al. Social determinants of breast cancer screening in urban primary care practices: a community-engaged formative study. Women's Health Issues. 2012; 22:e429-e38 [↑](#endnote-ref-54)
55. See <https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/2024-cancer-facts-figures.html> Accessed on 2 3 2024. [↑](#endnote-ref-55)
56. Zhang S, Xu H, Zhang L, Qiao Y. Cervical cancer: epidemiology, risk factors and screening. Chin J Cancer Res. 2020 Dec 31;32(6):720-728. [↑](#endnote-ref-56)
57. PDQ Cancer Information Summaries [Internet]. Bethesda (MD): National Cancer Institute (US); 2002-. Cervical cancer causes, risk factors, and prevention: patient version. 2023 Aug 18. Available from: https://www.ncbi.nlm.nih.gov/books/NBK65901/ [↑](#endnote-ref-57)
58. Fakhri N, Chad MA, Lahkim M, Houari A, Dehbi H, Belmouden A, El Kadmiri N. Risk factors for breast cancer in women: an update review. Med Oncol. 2022 Sep 7;39(12):197. [↑](#endnote-ref-58)
59. Sun YS, Zhao Z, Yang ZN, Xu F, Lu HJ, Zhu ZY, Shi W, Jiang J, Yao PP, Zhu HP. Risk factors and preventions of breast cancer. Int J Biol Sci. 2017 Nov 1;13(11):1387-1397. [↑](#endnote-ref-59)
60. de Groot P, Munden RF. Lung cancer epidemiology, risk factors, and prevention. Radiol Clin North Am. 2012 Sep;50(5):863-76. [↑](#endnote-ref-60)
61. Akhtar N, Bansal JG. Risk factors of Lung Cancer in nonsmoker. Current problems in cancer. 2017 Sep 1;41(5):328-39. [↑](#endnote-ref-61)
62. Lewandowska A, Rudzki G, Lewandowski T, Stryjkowska-Góra A, Rudzki S. Risk factors for the diagnosis of colorectal cancer. Cancer Control. 2022 Jan-Dec;29: 10732748211056692. [↑](#endnote-ref-62)
63. Vickers AJ, Cronin AM, et al. The relationship between prostate-specific antigen and prostate cancer risk: the Prostate Biopsy Collaborative Group. Clin Cancer Res. 2010 Sep 1;16(17):4374-81. [↑](#endnote-ref-63)
64. Fisher B, Costantino JP, et al. Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. J Natl Cancer Inst. 1998 Sep 16;90(18):1371-88. [↑](#endnote-ref-64)
65. Kattan MW, Eastham JA, et al. A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. J Natl Cancer Inst. 1998 May 20;90(10):766-71. [↑](#endnote-ref-65)
66. Peele PB, Siminoff LA, et al. Decreased use of adjuvant breast cancer therapy in a randomized controlled trial of a decision aid with individualized risk information. Med Decis Making. 2005 May-Jun;25(3):301-7. [↑](#endnote-ref-66)
67. Gill S, Loprinzi CL, et al. Pooled analysis of fluorouracil-based adjuvant therapy for stage II and III colon cancer: who benefits and by how much? J Clin Oncol. 2004 May 15;22(10):1797-806. [↑](#endnote-ref-67)
68. Núñez ER, Lindenauer PK, Wiener RS. Electronic Health Record-Based Algorithms as Part of the Solution for Improving Lung Cancer Screening. JCO Clin Cancer Inform. 2023 Sep;7:e2300222. [↑](#endnote-ref-68)
69. Callender T, Imrie F, et al. Assessing eligibility for lung cancer screening using parsimonious ensemble machine learning models: A development and validation study. PLoS Med. 2023 Oct 3;20(10):e1004287. [↑](#endnote-ref-69)
70. Burnett-Hartman AN, Powers JD, et al. Development of an electronic health record-based algorithm for predicting lung cancer screening eligibility in the population-based research to optimize the screening process Lung Research Consortium. JCO Clin Cancer Inform. 2023 Sep;7:e2300063. [↑](#endnote-ref-70)
71. Nguyen QTN, Phan PT, et al. Machine-learning based risk assessment for cancer therapy-related cardiac adverse events among breast cancer patients. Stud Health Technol Inform. 2024 Jan 25;310:1006-1010. [↑](#endnote-ref-71)
72. Louro J, Román M, et al. Personalized breast cancer screening: A risk prediction model based on women attending breast screen Norway. Cancers (Basel). 2023 Sep 12;15(18):4517. [↑](#endnote-ref-72)
73. Deng Y, Lu Y, et al. Prediction of human epidermal growth factor receptor 2 (HER2) status in breast cancer by mammographic radiomics features and clinical characteristics. Eur Radiol. 2024 Jan 26. [↑](#endnote-ref-73)
74. Munshi RM. Novel ensemble learning approach with SVM-imputed ADASYN features for enhanced cervical cancer prediction. PLoS One. 2024 Jan 10;19(1):e0296107. [↑](#endnote-ref-74)
75. Das K, Paltani M, et al. Current implications and challenges of artificial intelligence technologies in therapeutic intervention of colorectal cancer. Explor Target Antitumor Ther. 2023;4(6):1286-1300. [↑](#endnote-ref-75)
76. Halm EA, Nair RG, et al. Improving Colorectal Cancer Screening in a Regional Safety-Net Health System over a 10-Year Period: Lessons for Population Health. J Gen Intern Med. 2023 Nov 6. [↑](#endnote-ref-76)
77. Petrik AF, Johnson ES, et al. The use of individual and multilevel data in the development of a risk prediction model to predict patients' likelihood of completing colorectal cancer screening. Prev Med Rep. 2023 Aug 11;36:102366. [↑](#endnote-ref-77)
78. El Haji H, Souadka A, et al. Evolution of breast cancer recurrence risk prediction: A systematic review of statistical and machine learning–based models. JCO Clinical Cancer Informatics. 2023;7 :e2300049. [↑](#endnote-ref-78)
79. Mertens E, Barrenechea-Pulache A, et al. Understanding the contribution of lifestyle in breast cancer risk prediction: a systematic review of models applicable to Europe. BMC cancer. 2023 Jul 21;23(1):687. [↑](#endnote-ref-79)
80. McBride KA, Ballinger ML, et al. Li-Fraumeni syndrome: cancer risk assessment and clinical management. Nat Rev Clin Oncol. 2014 May;11(5):260-71. [↑](#endnote-ref-80)
81. Ferreira CR. The burden of rare diseases. Am J Med Genet A. 2019 Jun;179(6):885-892. [↑](#endnote-ref-81)
82. Blazsik RM, Beeler PE, et al. Impact of single and combined rare diseases on adult inpatient outcomes: a retrospective, cross-sectional study of a large inpatient population. Orphanet J Rare Dis. 2021 Feb 27;16(1):105. [↑](#endnote-ref-82)
83. Young RP, Hopkins RJ. How the genetics of lung cancer may overlap with COPD. Respirology. 2011 Oct;16(7):1047-55. [↑](#endnote-ref-83)
84. Ren M, Lu C, Zhou M, Jiang X, Li X, Liu N. The intersection of virus infection and liver disease: A comprehensive review of pathogenesis, diagnosis, and treatment. WIREs Mech Dis. 2024 Jan 22:e1640. [↑](#endnote-ref-84)
85. Schiller JT, Lowy DR. An Introduction to virus infections and human cancer. Recent Results Cancer Res. 2021;217:1-11. [↑](#endnote-ref-85)
86. Tu H, Wen CP, et al. Cancer risk associated with chronic diseases and disease markers: prospective cohort study. BMJ. 2018 Jan 31;360:k134. [↑](#endnote-ref-86)
87. Yan F, Eshak ES, Arafa A, Tamakoshi A, Iso H; JACC Study Group. Vitamin K intake and risk of lung cancer: The Japan Collaborative Cohort Study. J Epidemiol. 2023 Oct 5;33(10):536-542. [↑](#endnote-ref-87)
88. Ferdosnejad K, Zamani MS, Soroush E, Fateh A, Siadat SD, Tarashi S. Tuberculosis and lung cancer: metabolic pathways play a key role. Nucleosides Nucleotides Nucleic Acids. 2024 Feb 2:1-20. [↑](#endnote-ref-88)
89. Ten HK. Risk-based lung cancer screening eligibility criteria: towards implementation. The Lancet Oncology. 2021. [↑](#endnote-ref-89)
90. Investigators AoURP. The “All of Us” research program. New England Journal of Medicine. 2019; 381:668-76. [↑](#endnote-ref-90)
91. See https://databrowser.researchallofus.org/ehr/conditions/cancer accessed on 2/2/2024. [↑](#endnote-ref-91)
92. Aschebrook-Kilfoy B, Zakin P, et al. An overview of cancer in the first 315,000 All of Us participants. PLoS One. 2022 Sep 1;17(9): e0272522. [↑](#endnote-ref-92)
93. Maksut J, Hodge C, et al. Utilization of Z codes for social determinants of health among medicare fee-for-service beneficiaries, 2019. Centers for Medicare and Medicaid Services (CMS). 2021. [↑](#endnote-ref-93)
94. Truong HP, Luke AA, et al. Utilization of social determinants of health ICD-10 Z-codes among hospitalized patients in the United States, 2016-2017. Medical care. 2020; 58:1037. [↑](#endnote-ref-94)
95. Ubri PS, Bailey L, et al. The Role of State Policy in Use of Z Codes to Document Social Need in Medicaid Data. 2022. [↑](#endnote-ref-95)
96. Association AH. ICD-10-CM coding for social determinants of health. American Hospital Association, Chicago, IL; 2020. [↑](#endnote-ref-96)
97. Alemi F, Avramovic S, Renshaw KD, Kanchi R, Schwartz M. Relative accuracy of social and medical determinants of suicide in electronic health records. Health services research. 2020; 55:833-40 [↑](#endnote-ref-97)
98. Glymour M, Pearl J, Jewell NP. Causal inference in statistics: A primer: John Wiley & Sons; 2016. [↑](#endnote-ref-98)
99. Pearl J. Causality: Models, reasoning and inference. Cambridge, UK: Cambridge University Press. 2000; 19. [↑](#endnote-ref-99)
100. Garcea F, Morra L, Lamberti F. On the Use of Causal Models to Build Better Datasets. In2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC) 2021 Jul 12 (pp. 1514-1519). IEEE. [↑](#endnote-ref-100)
101. Alemi F. Worry less about the algorithm, more about the sequence of events. Math Biosci Eng. 2020 Sep 27;17(6):6557-6572. [↑](#endnote-ref-101)
102. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983, 70, 41–55. [↑](#endnote-ref-102)
103. Pearl J. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2000. [↑](#endnote-ref-103)
104. Alemi F, ElRafey A, Avramovic I. Covariate balancing through naturally occurring strata. Health Serv Res. 2018; 53(1): 273-292. [↑](#endnote-ref-104)
105. Gendron G, Witbrock M, Dobbie G. A Survey of methods, challenges and perspectives in causality. arXiv preprint arXiv:2302.00293. 2023 Feb 1. [↑](#endnote-ref-105)
106. Seth AK. Causal networks in simulated neural systems. Cognitive neurodynamics. 2008 Mar;2(1):49-64. [↑](#endnote-ref-106)
107. Ma S, Liu L, Li J, Le TD. Data-driven discovery of causal interactions. International Journal of Data Science and Analytics. 2019 Oct;8:285-97. [↑](#endnote-ref-107)
108. Alemi F. Constructing causal networks through regressions: A tutorial. Quality Management in Health Care. 2020; 29:270-8. [↑](#endnote-ref-108)
109. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: Springer; 2013. [↑](#endnote-ref-109)
110. Wu C, Ma S. A selective review of robust variable selection with applications in bioinformatics. Brief Bioinform 2015. Sep;16(5):873-83. [↑](#endnote-ref-110)
111. Musoro JZ, Zwinderman AH, Puhan MA, ter Riet G, Geskus RB. Validation of prediction models based on lasso regression with multiply imputed data. BMC Med Res Methodol. 2014 Oct 16;14:116. [↑](#endnote-ref-111)
112. Brier GW. Verification of forecasts expressed in terms of probability. Monthly Weather Review 1950. 78: 1–3. [↑](#endnote-ref-112)
113. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007; 115: 928-35. [↑](#endnote-ref-113)
114. Pencina MJ, D'Agostino RB, et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008; 27: 157-72. [↑](#endnote-ref-114)
115. Glueck DH, Muller KE. Adjusting power for a baseline covariate in linear models. Statistics in medicine. 2003; 22:2535-51. [↑](#endnote-ref-115)
116. Muller KE, Lavange LM, et al. Power calculations for general linear multivariate models including repeated measures applications. Journal of the American Statistical Association. 1992; 87:1209-26. [↑](#endnote-ref-116)
117. Finch WH, Finch MEH. Regularization methods for fitting linear models with small sample sizes: Fitting the lasso estimator using R. Practical Assessment, Research, and Evaluation. 2016; 21:7 [↑](#endnote-ref-117)
118. Ramirez AH, Sulieman L, et al. All of Us Research Program. The All of Us Research Program: Data quality, utility, and diversity. Patterns (N Y). 2022 Aug 12;3(8):100570. [↑](#endnote-ref-118)
119. O’Brien KM, Mooney T, et al. Screening status, tumour subtype, and breast cancer survival: A national population-based analysis. Breast Cancer Research and Treatment, 2018, 172(1), 133–142. [↑](#endnote-ref-119)
120. Fonga Y, Evansb J, et al. The incidence and 10-year survival of interval breast cancers in Wales. Clinical Radiology 2014: 69, e168–172. [↑](#endnote-ref-120)
121. Hamashima C, Shabana M, Okamoto M, Osaki Y, Kishimoto T. Survival analysis of patients with interval cancer undergoing gastric cancer screening by endoscopy. PLoS One. 2015 May 29;10(5):e0126796. [↑](#endnote-ref-121)
122. Irvin VL, Zhang Z, et al. Comparison of mortality among participants of women’s health initiative trials with screening-detected breast cancers vs interval breast cancers. JAMA Network Open 2020, 3(6), e207227. [↑](#endnote-ref-122)
123. Su PF, Lin CK, Hung JY, Lee JS. The proper use and reporting of survival analysis and Cox regression. World Neurosurg. 2022 May;161:303-309. [↑](#endnote-ref-123)
124. Li W, Chen SY, Rong A. Estimation of delay time in survival data with delayed treatment effect. Journal of Biopharmaceutical Statistics. 2019 Mar 4;29(2):229-43. [↑](#endnote-ref-124)
125. Sit T, Liu M, Shnaidman M, Ying Z. Design and analysis of clinical trials in the presence of delayed treatment effect. Statistics in Medicine. 2016 May 20;35(11):1774-9. [↑](#endnote-ref-125)
126. Flandre P, O’Quigley J. Comparing Kaplan–Meier curves with delayed treatment effects: applications in immunotherapy trials. Journal of the Royal Statistical Society Series C: Applied Statistics. 2019; 68(4):915-39. [↑](#endnote-ref-126)
127. Wyatt RM, Beddoe AH, Dale RG. The effects of delays in radiotherapy treatment on tumour control. Physics in Medicine & Biology. 2002 Dec 23;48(2):139 [↑](#endnote-ref-127)
128. Flandre P, O’Quigley J. Comparing Kaplan–Meier curves with delayed treatment effects: applications in immunotherapy trials. Journal of the Royal Statistical Society Series C: Applied Statistics. 2019; 68(4):915-39 [↑](#endnote-ref-128)
129. Krist AH, Davidson KW, Mangione CM, Barry MJ, Cabana M, Caughey AB, et al. Screening for lung cancer: US Preventive Services Task Force recommendation statement. Jama. 2021; 325:962-70. [↑](#endnote-ref-129)
130. Kukhareva PV, Caverly TJ, et al. Inaccuracies in electronic health records smoking data and a potential approach to address resulting underestimation in determining lung cancer screening eligibility. J Am Med Inform Assoc. 2022 Apr 13;29(5):779-788. [↑](#endnote-ref-130)
131. Liu S, McCoy AB, et al. Leveraging natural language processing to identify eligible lung cancer screening patients with the electronic health record. Int J Med Inform. 2023 Sep;177:105136. [↑](#endnote-ref-131)
132. Kats DJ, Adie Y, et al. Assessing Different Approaches to Leveraging Historical Smoking Exposure Data to Better Select Lung Cancer Screening Candidates. Nicotine Tob Res. 2021 Aug 4;23(8):1334-1340. [↑](#endnote-ref-132)
133. López-Ratón M, Rodríguez-Álvarez MX, et al. OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. Journal of Statistical Software. 2014; 61(8): 36. [↑](#endnote-ref-133)
134. Patel AV, Deubler E, et al. Key risk factors for the relative and absolute 5-year risk of cancer to enhance cancer screening and prevention. Cancer 2022, 128(19), 3502–3515. [↑](#endnote-ref-134)
135. Komoto S, Nishiwaki Y, et al. Effect size and cost-effectiveness estimates of breast and cervical cancer screening reminders by population size through complete enumeration of Japanese local municipalities. BMC Public Health 2014, 14, 43. [↑](#endnote-ref-135)
136. Cheung LC, Ramadas K, et al. Risk-Based Selection of Individuals for Oral Cancer Screening. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2021, 39(6), 663–674. [↑](#endnote-ref-136)
137. Eapen RS, Herlemann A, et al. Impact of the United States Preventive Services Task Force 'D' recommendation on prostate cancer screening and staging. Curr Opin Urol. 2017 May;27(3):205-209. [↑](#endnote-ref-137)
138. Plambeck BD, Wang LL, et al. Effects of the 2012 and 2018 US preventive services task force prostate cancer screening guidelines on pathologic outcomes after prostatectomy. Prostate. 2022 Feb;82(2):216-220. [↑](#endnote-ref-138)
139. Maki KG, Talluri R, et al. Impact of U.S. Preventive Services Task Force lung cancer screening update on drivers of disparities in screening eligibility. Cancer Med. 2023 Feb; 12(4): 4647-4654. [↑](#endnote-ref-139)
140. Fleshner K, Carlsson SV, Roobol MJ. The effect of the USPSTF PSA screening recommendation on prostate cancer incidence patterns in the USA. Nat Rev Urol. 2017 Jan;14(1):26-37. [↑](#endnote-ref-140)
141. Soori M, Platz EA, Brawley OW, Lawrence RS, Kanarek NF. Inclusion of the US Preventive Services Task Force Recommendation for Mammography in State Comprehensive Cancer Control Plans in the US. JAMA Netw Open. 2022 May 2;5(5):e229706. [↑](#endnote-ref-141)