Supplementary: Linear Cancer Prediction Models with Novel Features Outperform Previous Approaches

Scott Kulm[1,2,3], Lior Kofman[2], Jason Mezey[4,5,*], and Olivier Elemento[2,3,*]

[1]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY
[2]Caryl and Israel Englander Institute of Precision Medicine, Weill Cornell Medicine, New York, NY
[3]Physiology, Biophysics and Systems Biology Graduate Program, Weill Cornell Medicine, New York, NY
[4]Department of Genetic Medicine, Weill Cornell Medicine, New York, NY
[5]Department of Computational Biology, Cornell University, Ithaca, NY
[*]corresponding authors

Online Methods

The data used in this study was derived from the UK Biobank. First, the UK Biobank was manually filtered to ethnically white British individual and six classes of features: PGS, EHR, lifestyle, medical, census and biomarkers. The filtering of individuals was necessary as the polygenic risk scores have been derived and validated within Europeans and extrapolation to other ethnicities is likely to have significantly decreased efficacy. Genetic data was determined through genotyping with either the BiLEVe or Axiom chip, followed by imputation with the IMPUTE2 program. The polygenic risk scores were then constructed using the listed single nucleotide polymorphisms and associated weights in the Graff et al. publication within the PLINK software. The ICD features were extracted from reported hospital intake data (HESIN), with a binary ICD feature confirmed if the respective ICD code was reported anytime before the individual was initially assessed by the UK Biobank. The ICD features were filtered to a minimum of 500 events, and any relevant cancer ICD code was removed for future consideration (ICD Z08, ICD Z85, ICD C77, OPCS X72, OPCS X70). The lifestyle and medical category derives from survey answers which were the responses given during the initial assessment. The lifestyle category includes questions that align more closely with life contentment, diet, exercise and hobbies whereas the medical category includes questions that align more closely with measurements and health history. Some features could likely be classified in either category. The census data was obtained by linking the reported home location of each individual (accurate to within 1 km) to the individual's local super output area. The census data for that local super output area, obtained from the United Kingdom Office for National Statistics, was then ascribed to that individual and that respective census feature. Lastly, biomarkers included antibody titers, blood cell counts, and other common compounds recorded in a blood screen. Please see feature spreadsheets within the supplementary data for a full list of features considered. In total 931 features and 408,166 individuals were included.

Cancer outcomes were called using the Surveillance, Epidemiology, and End Results Program (SEER) recodes. Similar to how ICD features were called, if an individual reported any of the respective SEER ICD-10 codes (Supplementary Table 2) from the time of initial assessment to 10 years time from the initial assessment then the individual was considered to be positive for that cancer. If the individual was positive for a cancer before the initial assessment time they were removed from future analyses for that specific cancer. We additionally considered individuals to be positive for a previous cancer if they reported specific other codes not included in the SEER recodes. We took this step to ensure that individuals with a recurrent diagnosis would not be included in our investigation. Please note that all predictive features were thereby recorded before the initial assessment, and all cancer outcomes were considered only after the initial assessment.

Quality control of the predictive features started by removing features that contained high proportions of missing data. Next, categorical features were transformed such that each category became its own binary feature. Only non-missing entries in the categorical features led to non-zero entries in one of the generated binary features. Any feature whose values were over 75% missing and any binary feature (including those generated from categorical features) whose minor class included fewer than 500 values were removed. Next, all features were standardized such that they had a standard range from zero to one. Then pairwise spearman correlation was computed across all features, except for pairs of binary features, for which a F1 score was computed. For each pair of features that had a correlation or F1 score greater than 0.95 one feature was randomly chosen and removed from the larger feature set. Next, the data was randomly split by a proportion of 70:30 intro training and testing sets. Following, all features in each data set were imputed using the iterativeImputer function within the sklearn library. To reduce time of computation each category of features were imputed separately, along with a specified set of salient features. The training and testing data was finally saved and used directly as the predictive feature set in all computations.

With the clean feature set and cancer diagnoses, predictive models were fit. First, the best machine learning model was sought out of XGBoost, AdaBoost, and Random Forest. As each of these models are sensitive to the features and parameters employed, an exhaustive cross-validation process attempted to find the best possible form of each machine learning model. The model selection process began with all available features. A genetic algorithm was then employed (Figure 1), which created a population of 10 models with hyper-parameters randomly chosen from a pre-described set. Over the course of 10 generations the two models that performed best were retained along with one other random model. The remaining seven models were formed by "breeding" or mixing and matching the hyper-parameters from two other models, followed by "mutating" or randomly changing one hyper-parameter. The best model throughout this process was defined by the AUC computed through three-fold cross validation (Figure 2), repeated three times. Once the genetic algorithm was complete the feature importances from the best model were extracted and used to subset the larger feature set (Feature 3). The genetic algorithm was then again employed. This process was repeated, with features limited to the 500, 100, 50, and 10 most important. The model with the highest AUC across all possible hyper-parameters, features, and algorithms was selected as the sole machine learning model to be fit on the entire training data-set and assessed on the testing data-set. Although, if the machine learning model chosen was not XGBoost, then the best performing XGBoost model would similarly be fit on the entire training data-set for the single purpose of extracting feature gains and interactions. The feature importances from the possibly alternative model were also retained and compared. An extended explanation of genetic algorithms is provided at `https://www.sciencedirect.com/topics/engineering/genetic-algorithm`.

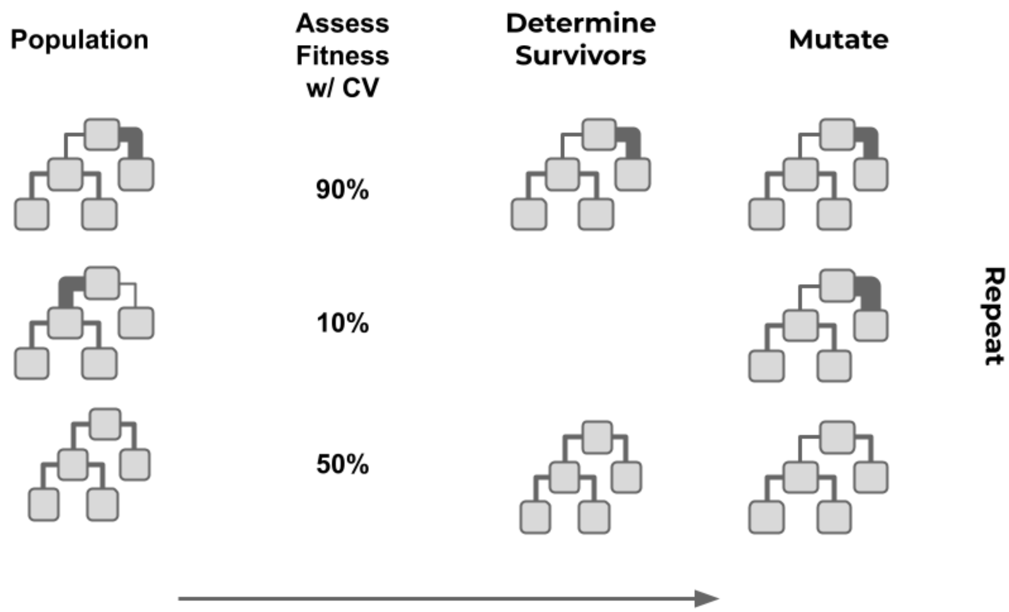

Figure 1. Flow chart describing the genetic algorithm. A population of models, each of which contains different hyperparameters, are assessed. Their fitness, or accuracy, determines which models survives to the next generation. Some of the remaining models are mutated and some models that have relatively low fitness are also retained. This process is repeated until the algorithm converges to a set of models with high-performing hyperparameters.

In addition to the machine learning models, linear models were formed with elastic net logistic regression. Specifically, two linear models were formed. The first was fit with the lambda value that maximized the accuracy during 10 fold cross validation within the training set (Figure 2 and 4). The second was fit with the lambda value that limited the number of resulting features to 10. Models with ten features, as opposed to nine or eleven, were chosen because the median number of features in the available QCancer models was ten, and the additional of other limited-feature linear models would likely overfit the testing data-set. Also, from prior data exploration we do not believe that there would be a considerable drop in accuracy if a nine or eleven feature model was used, so we considered the solitary use of the ten feature model to be preferred. An important note to make is that while we attempted to only include ten features in each ten feature model, on some occasions the algorithm would not conform and would include nine or eleven features in the model. In both instances the alpha was set to 1. We would like to highlight that this feature selection and modeling fitting process was completely unbiased as we did nothing to prefer any of the 961 features over another. The ten features that were retained in the ten feature linear model were chosen because of the computational algorithm. While this process has no guarantee of selecting features with a causal relationship, or those that have been previously validated, they are statically the collection of features that lead to the best possible predictions according to a robust, cross-validation procedure. The R package glmnet was used to fit the models and specifically the cv.glmnet was used to find the lambda value for the first model. Additional information on the elastic net process can be found at `https://hastie.su.domains/Papers/elasticnet.pdf`.

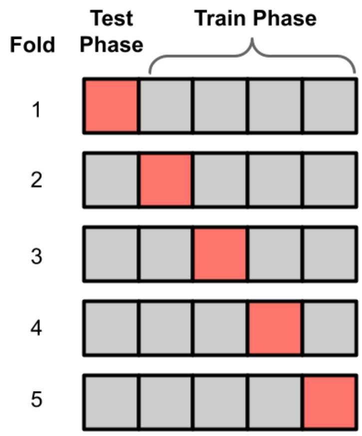

Figure 2. Flow chart describing the cross validation process. The total population is split into five folds, within each fold a different fifth of the population is withheld for evaluation or testing while remaining four fifths is used for training the model. This process can be used in conjunction with genetic algorithms to determine the accuracy of a set of features, or with an elastic net algorithm to determine the accuracy that corresponds to a lambda value.

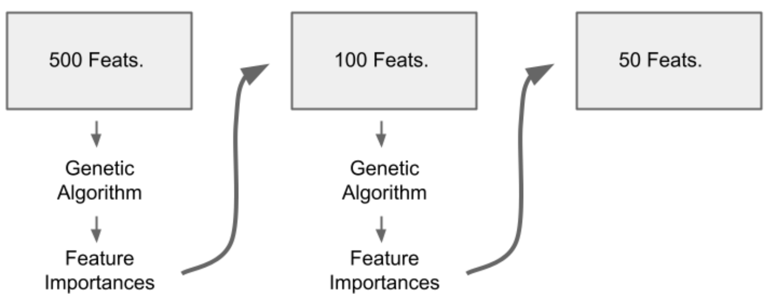

Figure 3. The feature selection process applied to the machine learning models. In each iteration a set of features is used to fit and assess models with various hyperparameters in a cross validation framework. The model that generates the greatest AUC is used to determine the importance of each feature. These importances are used to rank and subset the features into a new feature set in which the process repeats.

The last type of model formed were traditional linear models termed QCancer. Formed from QResearch, which is a non-profit collaboration between the University of Oxford and EMIS that holds patient records for 35 million patients, the QCancer models have been carefully derived and validated. The models are freely available for download and can be executed with relatively basic patient information. The resulting risk predictions were directly compared with the cancer diagnoses previously prepared.

The predictions for each cancer model, and the underlying features, were finally comprehensively analyzed. First, the features themselves were compared between models by either comparing the linear model coefficients, machine learning model's feature importances or the XGBoost model's expected gain. The linear model coefficients were derived by extracting the feature names from the elastic net model and then fitting a normal logistic regression with only those features available. This added step enabled us to calculate standard errors and compute Wald tests for each feature. To make a fair comparison the absolute value of the feature's importance was divided by the sum of the absolute value of all features' importance. The resulting value indicated the proportional weight of the feature on the mode's prediction. The interactions determined by the

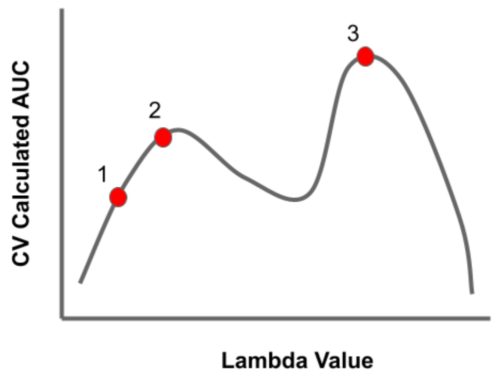

Figure 4. Process applied to determine the optimal lambda value of the elastic net algorithm. Various lambda values are applied within a cross validation framework, resulting in an AUC estiamte. This process is repeated across several pre-established lambda guesses. The lambda value that generates the highest AUC is retained for use in predicting the testing phase.

XGB model were verified through a series of linear models that included both interaction terms separately along with the interaction.

Second, the predictive performance of all models were compared. The primary methods used to determine predictive performance was through the construction of ROC curves and odds ratios. The ROC curves led to computation of AUC, and the point on the ROC curve in which 1 - TPR - FPR was at a minimum led to the reporting of accuracy, TPR and FPR values. Please note that these values can be easily modified to preference a higher TPR or FPR. The TPR refers to the number of true positives divided by the number of true positives plus false negatives, the FPR refers to the number of false positives divided by the number of false positives plus true negatives, and the accuracy refers to the number of true positives plus true negatives divided by the sum of all positives and all negatives. The odds ratios were calculated by designating individuals as being either exposed or non-exposed according to their risk and then constructing a typical 2x2 contingency table. Confidence intervals were constructed for AUC values through the De-Long method, and for odds ratios through the Wald approximation. Comparisons between predictions were either made between specific models through ROC curve comparisons through bootstrapping, or groups of cancers through paired Wilcoxon Rank-Sum tests.

Various factors influence the prediction accuracy of a specific model for a specific cancer. First, cancers with a greater number of cases are likely easier to predict owing to greater statistical power. Second, cancers that are caused by a single, high-impact feature are easy to predict as such a feature can be easily identified and incorporated into the model. Conversely, models that are caused by a disparate array of low-impact features are difficult to predict. Cancers that are caused by a large number of moderate-impact features, especially if those features interact with each other, are better predicted by more complex models as compared to simple models. These varying layers of high, moderate and low impact features can combine together in various ways, leading to a cancer that is generally easy to predict by some, but not all model types. Lastly, there are many unexplained causal factors that certainly influence the performance of a specific cancer by a specific model.

The programming language Python 3.6 was used to construct models and predictions, and R 3.6 was used for model analysis and plotting. The Python packages utilizes include: pandas, matplotlib, numpy, sklearn, xgboost, and xgbfir. The R packages utilized include: ggplot2, cowplot, stringr, reshape2, pROC, epitools, glmnet and data.table.

The UK Biobank Application 47137 was utilized in this investigation. All legal and ethical research approval concerning human subjects was obtained by the UK Biobank, and we have properly applied and been approved to work with their data-set.

Supplementary Figures and Tables

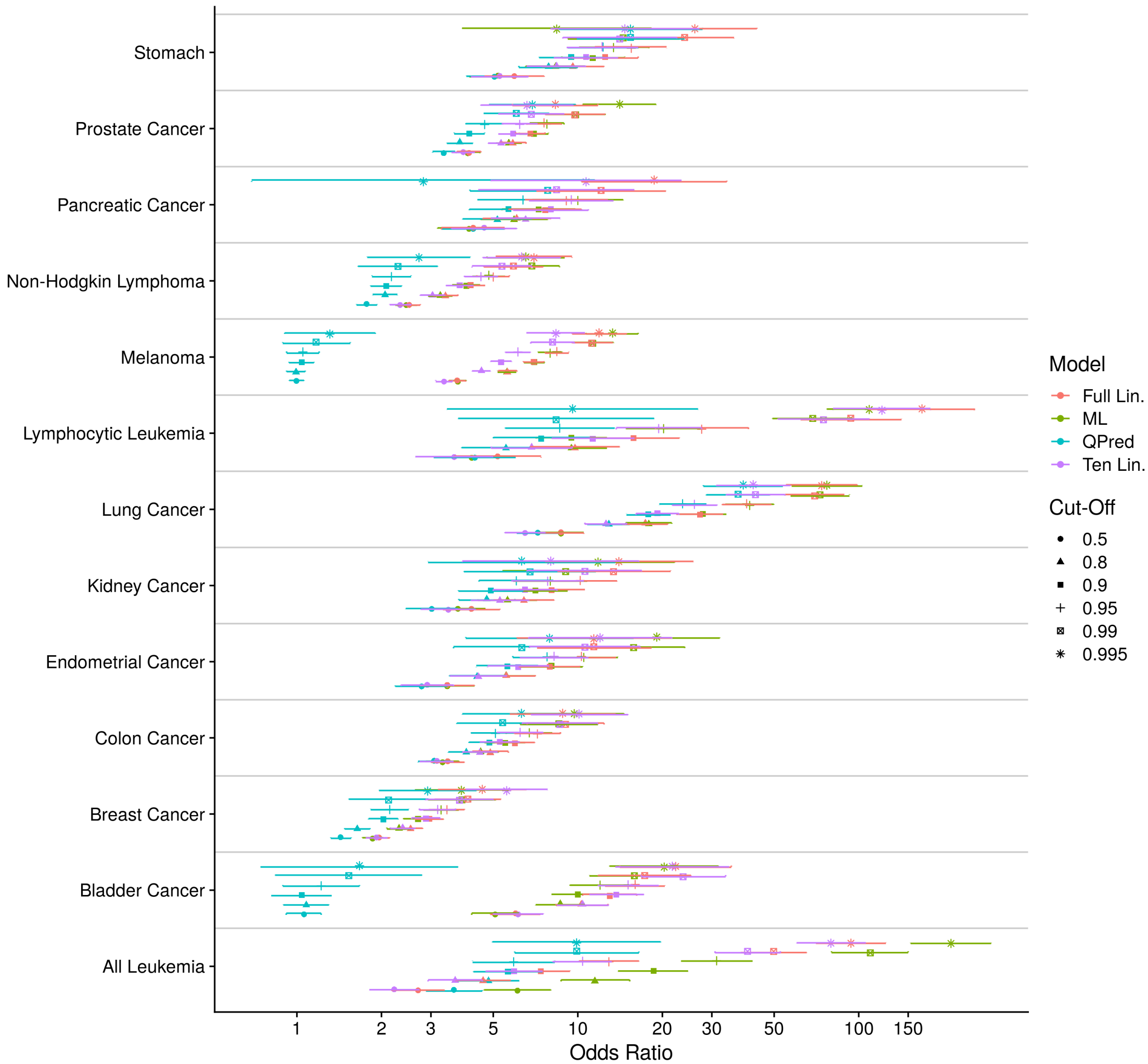

Figure 5. The odds ratios computed from various models and cut-offs, along with the respective 95% confidence intervals, for each cancer.

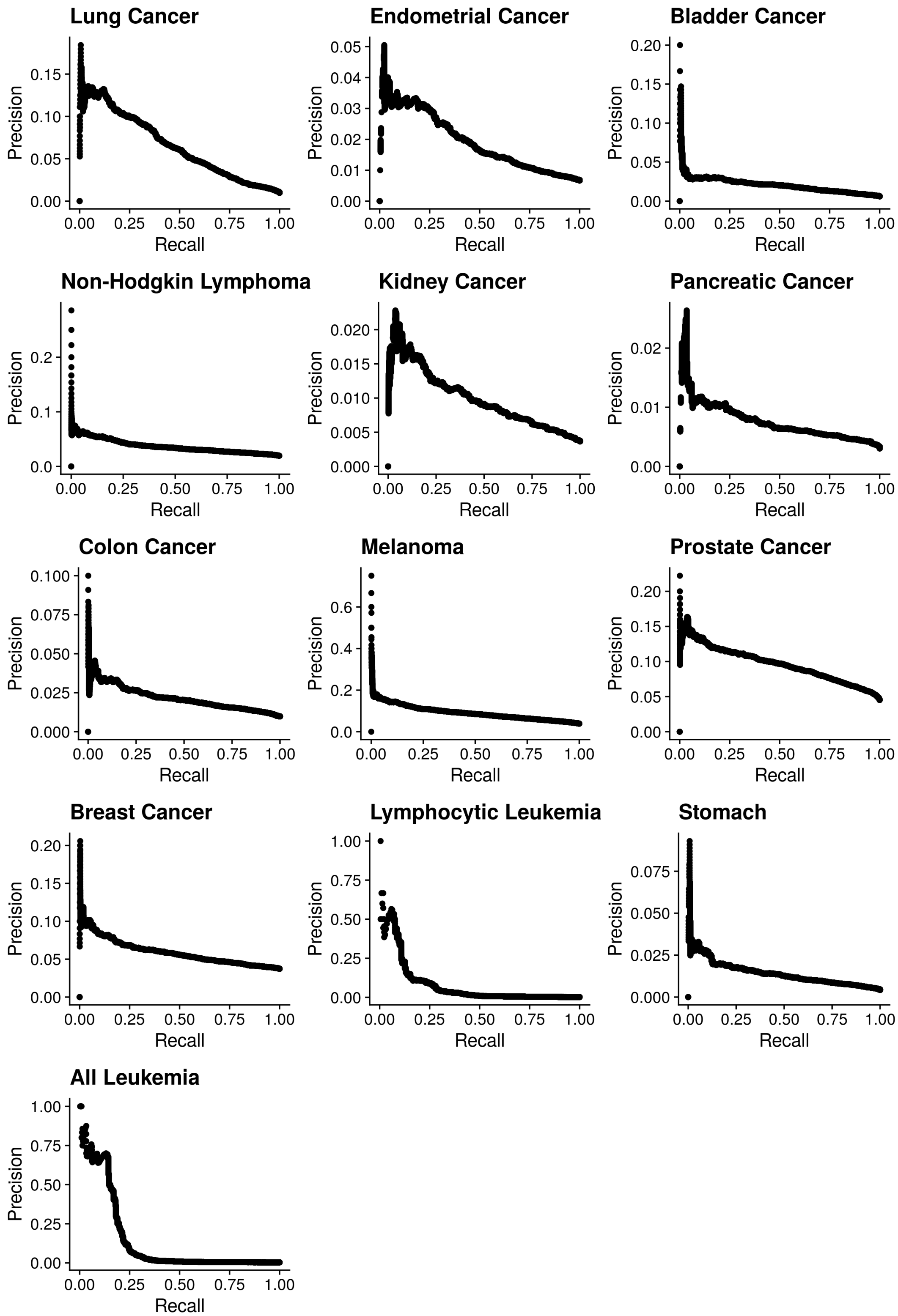

Figure 6. Precision-recall curves generated from full linear model predictions for each cancer.

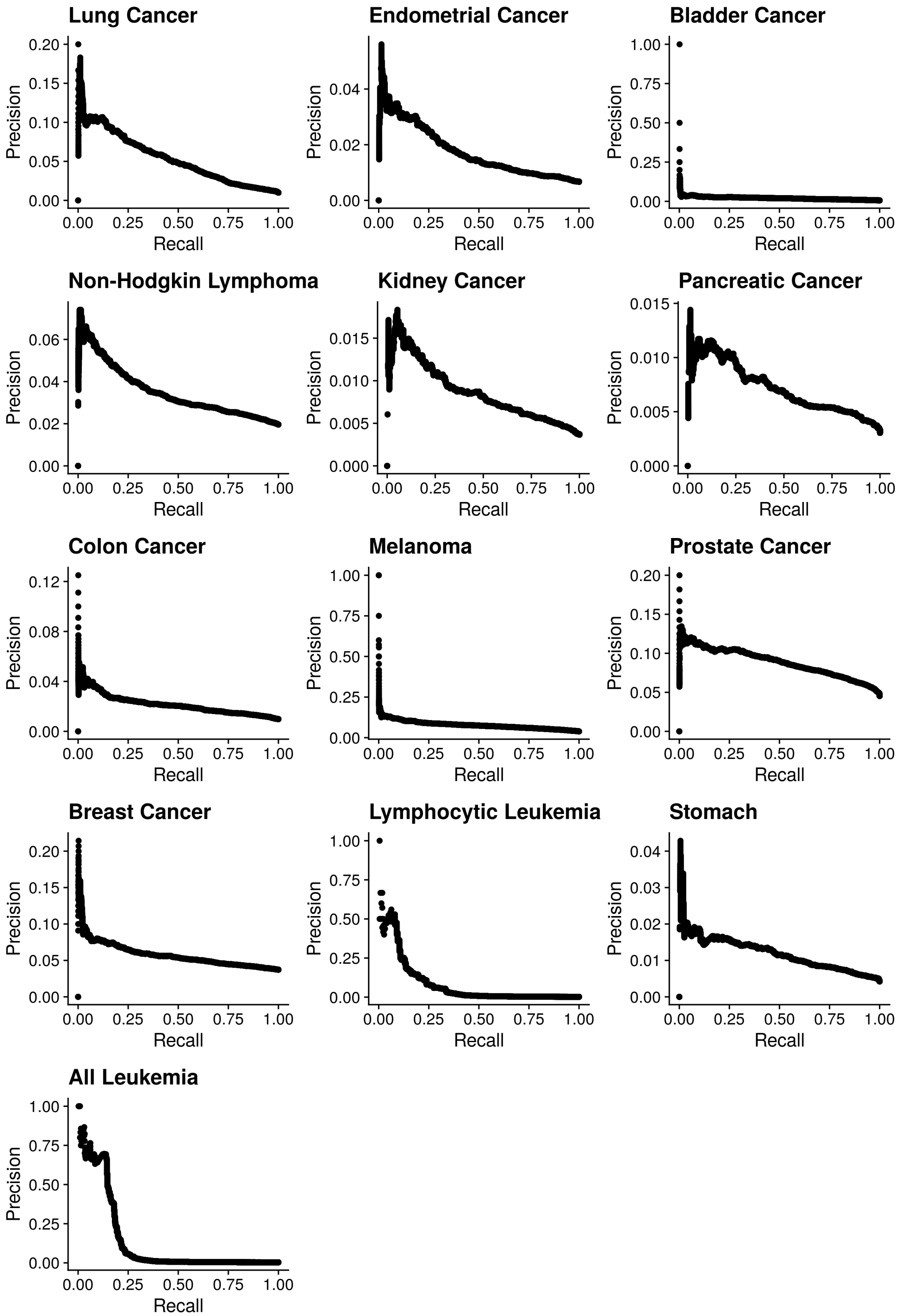

Figure 7. Precision-recall curves generated from ten-feature linear model predictions for each cancer.

Cancer	EHR Codes				
Lung Cancer					
Endometrial Cancer	ICD10_D06	ICD10_D05	ICD10_D26		
Bladder Cancer	ICD10_D30	ICD10_D41			
Oral	ICD10_D37	ICD10_D10			
Non-Hodgkin Lymphoma					
Kidney Cancer					
Ovarian Cancer	ICD10_D27				
Cervical Cancer	ICD10_D06				
Pancreatic Cancer	ICD10_D13				
Colon Cancer	ICD10_D13	ICD10_D12			
Melanoma	ICD10_D04	ICD10_D17	ICD10_D23		
Prostate Cancer					
Breast Cancer	ICD10_D05	ICD10_D24	OPCS_B27	OPCS_B28	MEDS_1140870164
Lymphocytic Leukemia					
Thyroid Cancer	ICD10_D34	ICD10_D35	ICD10_D44		
Testicular Cancer	ICD10_D01	ICD10_D13	ICD10_D37		
Stomach	ICD10_D13				
All Leukemia					

Table 1. Codes used to remove individuals from a cancer's respective analysis. If an individual reports any of the codes before the date of the assessment they would not be included in later model derivation or assessment.

Cancer	ICD-10 Code Definitions
Lung	C34 C340 C341 C342 C343 C348 C349
Endometrium	C54 C540 C541 C542 C543 C548 C549
Urinary	C67 C670 C671 C672 C673 C674 C675 C676 C677 C678 C679
Oral Cavity and Pharynx	C00 C03 C04 C05 C06 C09 C10 C11 C13 C000 C001 C002 C003 C004 C005 C006 C008 C009 C000 C001 C002 C003 C004 C005 C006 C008 C009 C000 C001 C002 C003 C004 C005 C006 C008 C009 C040 C041 C048 C049 C030 C031 C039 C050 C051 C052 C058 C059 C060 C061 C062 C068 C069 C110 C111 C112 C113 C118 C119 C090 C091 C098 C099 C100 C101 C102 C103 C104 C108 C109 C130 C131 C132 C138 C139 C140 C142
Non-Hodgkin's lymphoma	C024 C098 C099 C111 C142 C420 C421 C422 C770 C771 C772 C773 C774 C775 C778 C779 C424 C37
Kidney	C64 C65
Ovary	C56
Cervix	C53 C530 C531 C538 C539
Pancreas	C25 C250 C251 C252 C253 C254 C257 C258 C259
Colon and Rectum	C18 C180 C181 C182 C183 C184 C185 C186 C187 C188 C189 C260
Melanoma	C44 C440 C441 C442 C443 C444 C445 C446 C447 C448 C449
Prostate	C61
Breast	C50 C500 C501 C502 C503 C504 C505 C506 C508 C509
Lymphocytic Leukemia	C91 C910 C911 C912 C913 C914 C915 C916 C917 C918 C919
Thyroid	C73
Testis	C62 C620 C621 C629
Esophagus and Stomach	C15 C16 C150 C151 C152 C153 C154 C155 C158 C159 C160 C161 C162 C163 C164 C165 C166 C168 C169
All Leukemia	C91 C92 C93 C94 C95

Table 2. ICD-10 codes, taken from the SEER recodes, used to determine whether an individual was diagnosed with a given cancer. Specifically if any individual was given any of these codes from the time of assessment to the end of the study they would be considered a case in terms of that specific cancer, otherwise they would be a control.

	Lung Cancer		**Endo-metrial Cancer**		**Mela-noma**		**Prostate Cancer**	
	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
Sex								
Male	593	55459	0	0	2612	51987	2524	53196
Female	613	65777	436	64978	2008	62031	0	0
Year of Birth								
<1945	647	34781	181	17594	2281	31276	1233	15559
1946-1950	340	29158	112	15855	1257	27450	773	12486
1951-1960	193	36924	115	20506	832	35582	464	15907
>1960	26	20373	28	11023	250	19710	54	9244
Num. Self Rep. Illnesses								
0	161	30207	91	16197	896	28750	493	13362
1-2	502	56550	197	29956	2203	53224	1199	24969
>2	543	34468	148	18819	1518	32033	832	14856
Alcohol Use Freq.								
Never	124	7963	37	5158	265	7425	104	2549
Spec. Occ. Only	163	12871	88	9161	425	12029	140	3477
1-3 xMonth	127	13495	72	8422	429	12868	216	4717
1-2 xWeek	269	32129	97	17240	1220	30255	634	13979
3-4 xWeek	221	29095	82	13988	1168	27403	667	14504
Daily	299	25599	60	10972	1111	23954	761	13934
Smoking Status								
Never	205	66493	263	38412	2371	62339	1163	25992
Previous	537	42392	141	20711	1875	39631	1116	20734
Current	451	11917	30	5650	351	11680	233	6283
BMI								
<24	294	28462	75	19750	1070	27210	400	8490
24-27	325	35064	90	17883	1398	32925	798	16226
28-31	346	34434	98	15143	1359	32059	914	18129
>31	237	22753	166	11924	771	21340	404	10122

	Breast Cancer		**Lympho-cytic Leukemia**		**Stomach Cancer**		**All Leukemia**	
	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
Sex								
Male	0	0	150	55721	366	55995	265	55947
Female	2360	60938	72	66035	152	65953	159	65840
Year of Birth								
<1945	739	16295	106	34920	286	34871	221	35122
1946-1950	632	14719	75	29546	118	29334	118	29148
1951-1960	673	19447	33	36787	99	37260	71	37204
>1960	316	10477	8	20503	15	20483	14	20313
Num. Self Rep. Illnesses								
0	549	15330	40	30181	83	30665	81	30363
1-2	1113	28058	91	56851	222	56389	182	56790
>2	697	17545	91	34711	213	34884	161	34620
Alcohol Use Freq.								
Never	176	4880	11	7981	48	7986	28	7950
Spec. Occ. Only	304	8545	30	12937	60	12874	59	12995
1-3 xMonth	308	8078	14	13425	55	13468	42	13493
1-2 xWeek	620	16253	58	32266	136	32455	105	32164
3-4 xWeek	536	13049	55	29325	85	29279	86	29321
Daily	416	10084	54	25742	134	25806	103	25787
Smoking Status								
Never	1334	36217	92	66302	170	66375	195	66241
Previous	813	19226	104	42668	246	42778	180	42883
Current	204	5259	26	12367	96	12322	48	12259
BMI								
<24	622	18525	38	28965	75	28719	90	28747
24-27	649	16674	63	35037	143	35084	120	35129
28-31	597	14317	69	34254	173	34835	125	34567
>31	481	11146	47	22937	125	22755	87	22791

	Bladder Cancer		Non Hodgkin Lymphoma		Kidney Cancer		Pancreatic Cancer		Colon Cancer	
	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
Sex										
Male	555	54963	1027	54902	296	55927	198	55910	632	54711
Female	188	66264	1360	64435	153	66348	173	65597	553	64709
Year of Birth										
<1945	416	34857	962	33967	196	34909	188	34590	588	33684
1946-1950	193	29155	651	28551	142	29705	100	29515	321	28886
1951-1960	113	36783	601	36473	92	37306	73	37080	217	36538
>1960	21	20432	173	20346	19	20355	10	20322	59	20312
Num. Self Rep. Illnesses										
0	133	30161	513	29838	63	30552	71	30295	221	30262
1-2	336	56448	1098	55461	189	56694	163	56816	562	55637
>2	274	34603	776	34021	197	35020	136	34385	401	33509
Alcohol Use Freq.										
Never	44	7943	170	7887	31	7886	20	7839	96	7842
Spec. Occ. Only	71	12944	245	12690	61	13045	39	12723	141	12603
1-3 xMonth	63	13471	256	13303	55	13482	28	13523	93	13143
1-2 xWeek	172	32191	586	31640	129	32427	94	32231	286	31887
3-4 xWeek	165	29112	569	28727	95	29331	96	29446	269	28795
Daily	228	25485	559	25018	76	26028	94	25659	298	25073
Smoking Status										
Never	235	65867	1079	65418	191	66588	170	65952	554	65213
Previous	382	42654	945	41708	194	42948	151	42868	527	41785
Current	123	12250	357	11788	62	12313	48	12283	98	12031
BMI										
<24	112	28783	502	28264	71	28885	57	28534	216	28342
24-27	219	35011	675	34567	102	35128	107	35206	306	34444
28-31	247	34331	693	33993	139	34743	113	34821	370	34076
>31	158	22589	514	21999	136	23004	93	22432	287	22034

Table 3. Descriptive statistics enumerating the number of cases and controls in each row description.

Feature	Value
BMI	
Sex	Male
Year of Birth	
Smokers in Household	One household member
Smokers in Household	More than one household member
Smokers in Household	No
Smoking status	Never
Smoking status	Previous
Smoking status	Current
Alcohol Frequency	Daily
Alcohol Frequency	3 or 4 times a week
Alcohol Frequency	Once or twice a week
Alcohol Frequency	One to three times a month
Alcohol Frequency	Special Occasions
Alcohol Frequency	Never
Ever addicted to alcohol	Yes
General happiness	
General happiness with own health	
Ever addicted to illicit or recreational drugs	Yes
Ever addicted to prescription or over-the-counter medication	Yes
Cough on most days	Yes
Bring up phlegm/sputum/mucus on most days	Yes
Pack years of smoking	
Ever addicted to any substance or behaviour	Yes
Belief that own life is meaningful	Yes
Ever self-harmed	Yes
Age started smoking in former smokers	
Number of live births	

Table 4. The feature-set included during imputation of each other category of features. Imputation was carried out with the iterativeImputer function, and the single input to this function was a group of features that included both those listed in the table and a specific category (medical, lifestyle, EHR, census, biomarkers and PGS.

Cancer	ML	Full Linear	QCancer
Lung Cancer	494	242	16, 16
Endometrial Cancer	306	88	10
Bladder Cancer	50	94	
Non-Hodgkin Lymphoma	50	187	8,7
Kidney Cancer	100	68	12, 8
Pancreatic Cancer	10	44	9, 7
Colon Cancer	100	123	13, 14
Melanoma	50	276	
Prostate Cancer	100	113	10
Breast Cancer	100	196	14
Lymphocytic Leukemia	100	237	8, 7
Stomach	500	104	13, 11
All Leukemia	311	159	8, 7

Table 5. The number of features utilized for each model. The QCancer models may have two different number of features because there were one model for each sex.

Acronym	Meaning
Acc	Accuracy
ADA	AdaBoost
AUC	Area Under the Curve
EHR	Electronic Health Record
FPR	False Positive Rate
PGS	PolyGenic Risk Score
RF	Random Forest
ROC	Receiver Operating Characteristic
SEER	Surveillance, Epidemiology, and End Results Program
TPR	True Positive Rate
XGB	XGBoost

Table 6. Acronyms employed in the primary manuscript and their full meanings.

Cancer	ML	QCancer	Full.Lin.	Ten.Lin.
Lung Cancer	0.8334	0.7948	0.8282	0.8009
Endometrial Cancer	0.7117	NA	0.7153	0.6908
Bladder Cancer	0.7528	NA	0.7691	0.7682
Non-Hodgkin Lymphoma	0.6528	0.5975	0.6568	0.6402
Kidney Cancer	0.7119	0.6947	0.7236	0.7037
Pancreatic Cancer	0.7152	0.6895	0.723	0.7219
Colon Cancer	0.6894	0.6753	0.6949	0.6914
Melanoma	0.7099	NA	0.7103	0.6869
Prostate Cancer	0.7121	0.6765	0.7161	0.7055
Breast Cancer	0.6101	0.5632	0.6173	0.6102
Lymphocytic Leukemia	0.7166	0.7147	0.778	0.753
Stomach	0.7522	0.7407	0.7619	0.7466
All Leukemia	0.8033	0.6985	0.6945	0.6642

Table 7. The AUC for each model as assessed upon the testing data-set

Cancer	**XGB**	**RF**	**ADA**
Lung Cancer	0.8404	0.8367	0.8376
Endometrial Cancer	0.7222	0.7193	0.7214
Bladder Cancer	0.7833	0.7734	0.7811
Non-Hodgkin Lymphoma	0.6591	0.6522	0.6522
Kidney Cancer	0.7153	0.7189	0.7192
Pancreatic Cancer	0.7099	0.7042	0.715
Colon Cancer	0.6874	0.6844	0.6928
Melanoma	0.7114	0.7028	0.7026
Prostate Cancer	0.7156	0.7101	0.7086
Breast Cancer	0.6064	0.5986	0.6081
Lymphocytic Leukemia	0.8036	0.8045	0.802
Stomach	0.7672	0.761	0.7534
All Leukemia	0.7846	0.7819	0.7786

Table 8. The cross-validation derived AUC for each model that were then used to determine the sole machine learning model assessed on the testing data-set.

Cancer	ML:Full Lin.	ML:Ten Lin.	ML:QCancer	Full Lin.:Ten Lin.	Full Lin.:QCancer	Ten Lin.:QCancer
Lung Cancer	0.1371	2.36E-17	3.48E-25	1.89E-21	4.86E-12	0.2443
Endometrial Cancer	-0.6025	0.0042	1.32E-05	1.12E-07	8.45E-07	0.03532
Bladder Cancer	-0.000148	-4.92E-05	7.31E-72	0.6619	2.47E-87	1.39E-85
Non-Hodgkin Lymphoma	-0.2007	7.911E-07	7.903E-34	2.899e-09	5.674E-31	7.311E-23
Kidney Cancer	-0.0878	0.1583	0.03496	0.0002919	0.002513	0.327
Pancreatic Cancer	-0.3565	-0.3555	0.002497	0.7677	5.41E-05	1.56E-05
Colon Cancer	-0.1021	-0.3616	0.0008473	0.2419	1.59E-05	4.34E-05
Melanoma	-0.792	2.06E-39	2.216e-311	1.06E-28	8.533e-311	2.49E-245
Prostate Cancer	-0.09273	3.81E-05	4.56E-21	6.56E-08	3.05E-23	2.50E-15
Breast Cancer	-0.09615	-0.9745	4.97E-14	0.06565	3.39E-16	1.61E-14
Lymphocytic Leukemia	-0.002041	-0.06004	0.9286	0.009888	0.003079	0.0806
Stomach	-0.08146	0.3393	0.0338	0.00178	0.0007796	0.3075
All Leukemia	3.411e-15	5.214E-20	1.202E-23	3.018E-10	-0.8206	-0.06198

Table 9. The P-values extracted from De-Long's Test comparison of ROC curves. The ROC curves originated from assessment of the testing data-set with the two model predictions listed in the column header. The sign preceding the P-value indicates which model is superior, positive means the first model is superior and negative means the second model is superior.

Cancer	ML	Full.Lin.	Ten.Lin.	QCancer	Cut.Off
Lung Cancer	8.71	8.71	6.49	7.2	0.5
Lung Cancer	17.9	17.4	12.6	12.9	0.8
Lung Cancer	27.9	27.2	19.2	17.8	0.9
Lung Cancer	40.9	39.9	26	23.6	0.95
Lung Cancer	72.8	69.7	42.9	37.2	0.99
Lung Cancer	77	73.8	42.1	38.8	0.995
Endometrial Cancer	3.43	3.43	2.91	2.78	0.5
Endometrial Cancer	5.56	5.56	4.43	4.38	0.8
Endometrial Cancer	8.05	7.94	6.13	5.61	0.9
Endometrial Cancer	10.5	10.3	8.23	7.77	0.95
Endometrial Cancer	15.8	11.4	10.6	6.32	0.99
Endometrial Cancer	19.1	11.4	12	7.93	0.995
Bladder Cancer	5.08	6	6.13	0.981	0.5
Bladder Cancer	8.66	10.3	10.4	1.09	0.8
Bladder Cancer	10	13	13.7	1.13	0.9
Bladder Cancer	12	16	15.1	0.906	0.95
Bladder Cancer	15.9	17.3	23.7	0.665	0.99
Bladder Cancer	20.3	22.3	21.7	0.799	0.995
Non-Hodgkin Lymphoma	2.45	2.51	2.33	1.77	0.5
Non-Hodgkin Lymphoma	3.24	3.38	3.04	2.06	0.8
Non-Hodgkin Lymphoma	4.01	4.15	3.8	2.08	0.9
Non-Hodgkin Lymphoma	4.82	5	4.52	2.17	0.95
Non-Hodgkin Lymphoma	6.86	5.9	5.37	2.29	0.99
Non-Hodgkin Lymphoma	6.53	6.98	6.31	2.72	0.995
Kidney Cancer	3.74	4.18	3.46	3.02	0.5
Kidney Cancer	5.62	6.43	5.28	4.74	0.8
Kidney Cancer	7.07	8.07	6.49	4.9	0.9
Kidney Cancer	7.98	10.2	7.81	6.04	0.95
Kidney Cancer	9.05	13.4	10.6	6.76	0.99
Kidney Cancer	11.8	14	8.01	6.31	0.995
Pancreatic Cancer	4.1	4.24	4.64	4.24	0.5
Pancreatic Cancer	5.93	6.06	6.52	5.17	0.8
Pancreatic Cancer	7.25	7.66	8.01	5.66	0.9
Pancreatic Cancer	10	9.1	9.48	6.38	0.95
Pancreatic Cancer	NA	12.1	8.4	7.81	0.99
Pancreatic Cancer	NA	18.7	10.7	2.82	0.995
Colon Cancer	3.3	3.44	3.15	3.08	0.5
Colon Cancer	4.51	4.88	4.48	4.01	0.8
Colon Cancer	5.51	5.97	5.28	4.84	0.9
Colon Cancer	6.72	7.18	6.23	5.09	0.95
Colon Cancer	8.56	9.04	8.63	5.4	0.99
Colon Cancer	9.71	8.83	10.1	6.3	0.995
Melanoma	3.74	3.72	3.34	0.947	0.5
Melanoma	5.59	5.62	4.54	0.965	0.8
Melanoma	7	6.97	5.33	0.989	0.9
Melanoma	7.97	8.42	6.12	0.925	0.95
Melanoma	11.3	11.2	8.13	0.86	0.99
Melanoma	13.3	11.9	8.36	0.837	0.995
Prostate Cancer	4.06	4.1	3.91	3.33	0.5
Prostate Cancer	5.68	5.87	5.33	3.8	0.8
Prostate Cancer	6.98	6.78	5.88	4.11	0.9
Prostate Cancer	7.76	7.58	6.21	4.66	0.95
Prostate Cancer	9.82	9.75	6.83	6.04	0.99

Prostate Cancer	14.1	8.32	6.59	6.89	0.995
Breast Cancer	1.86	1.96	1.92	1.43	0.5
Breast Cancer	2.31	2.54	2.38	1.64	0.8
Breast Cancer	2.7	2.96	2.88	2.03	0.9
Breast Cancer	3.26	3.42	3.17	2.14	0.95
Breast Cancer	3.87	4.06	3.79	2.12	0.99
Breast Cancer	3.85	4.57	5.58	2.92	0.995
Lymphocytic Leukemia	4.19	5.18	3.63	4.3	0.5
Lymphocytic Leukemia	9.49	9.77	6.85	5.55	0.8
Lymphocytic Leukemia	9.49	15.8	11.3	7.41	0.9
Lymphocytic Leukemia	20.2	27.6	19.4	8.62	0.95
Lymphocytic Leukemia	68.6	93.7	74.9	8.37	0.99
Lymphocytic Leukemia	109	168	121	9.58	0.995
Stomach	5.2	5.94	5.27	5.05	0.5
Stomach	8.33	9.6	8.4	7.87	0.8
Stomach	11.3	12.5	10.7	9.47	0.9
Stomach	13.4	15.5	12.2	12.3	0.95
Stomach	14.5	24	14.1	15.4	0.99
Stomach	8.41	26.1	14.7	15.4	0.995
All Leukemia	6.1	2.7	2.22	3.62	0.5
All Leukemia	11.5	4.6	3.66	4.81	0.8
All Leukemia	18.6	7.39	5.93	5.64	0.9
All Leukemia	31.2	12.9	10.4	5.91	0.95
All Leukemia	110	49.8	40.2	9.91	0.99
All Leukemia	213	93.8	79.6	9.9	0.995

Table 10. The odds ratios computed for the machine learning, full linear, 10 feature linear and QCancer predictions. Odds ratios were computed through a contingency table in which the exposed or high-risk individuals have a prediction greater than the percentile of all predictions listed in the cut-off column and unexposed or low-risk individuals have a prediction less than the 50th percentile of all predictions. Therefore, the values listed are point estimates for a specific cancer, model and cut-off and are not averages.

Cancer	Feature	Type	Coefficient	P
Lung Cancer	White Blood Cell[1]	Biomarker	6.723	5.228e-06
Lung Cancer	Rent from Authority Lived In[2]	Lifestyle	0.5272	1.629e-09
Lung Cancer	Age[3]	Lifestyle	-3.396	6.161e-76
Lung Cancer	Number of treatments/medications taken[4]	Medical	1.097	0.01698
Lung Cancer	Smoking Status: Never[5]	Lifestyle	-0.9061	3.616e-26
Lung Cancer	Smoking Status: Current[5]	Lifestyle	1.105	2.714e-56
Lung Cancer	Pack Years of Smoking[5]	Lifestyle	5.887	3.088e-52
Lung Cancer	ICD: Problems Related to Lifestyle[6]	EHR	0.2888	0.007785
Lung Cancer	Medication: Tiotropium[7]	Medical	1.092	9.611e-08
Lung Cancer	OPCS: Operation on femoral artery[8]	EHR	0.6565	0.02177
Endometrial Cancer	Cyastin C[9]	Biomarker	-1.276	0.5896
Endometrial Cancer	BMI by Impedance[10]	Medical	5.329	9.433e-34
Endometrial Cancer	Had Menopause[11]	Medical	0.8562	2.163e-11
Endometrial Cancer	Ever taken oral contraceptive[12]	Lifestyle	-0.4849	8.33e-06
Endometrial Cancer	Age[3] Started Oral Contraceptive	Medical	1.129	0.04233
Endometrial Cancer	Smoking Status: Current[5]	Lifestyle	-0.1048	0.5853
Endometrial Cancer	PGS: Lymphocytic Leukemia[13]	PGS	1.784	9.552e-07
Endometrial Cancer	Medication: Tamoxifen[14]	Medical	1.086	0.002818
Endometrial Cancer	Medication: Digoxin[15]	Medical	-11.91	0.9611
Endometrial Cancer	OPCS: Abdominal excision of uterus[16]	EHR	-1.155	0.02207
Bladder Cancer	Cyastin C[9]	Biomarker	0.05583	0.9699
Bladder Cancer	HDL Cholesterol[17]	Biomarker	-0.138	0.7799
Bladder Cancer	Sex[18]	Medical	0.7546	2.101e-07
Bladder Cancer	Age[3]	Lifestyle	-3.453	1.84e-49
Bladder Cancer	Had Menopause[19]	Medical	-0.4346	0.007219
Bladder Cancer	Smoking Status: Never[5]	Lifestyle	-0.5133	5.097e-09
Bladder Cancer	Smoking Status: Current[5]	Lifestyle	0.2816	0.008676
Bladder Cancer	Pack Years of Smoking[5]	Lifestyle	2.554	3.072e-05
Bladder Cancer	PGS: Lung[13]	PGS	1.808	2.979e-07
Bladder Cancer	OPCS: Repair of recurrent inguinal hernia[20]	EHR	-0.2614	0.7146
Non-Hodgkin Lymphoma	Age[3]	Lifestyle	-1.683	6.329e-52
Non-Hodgkin Lymphoma	Smoking Status: Never[5]	Lifestyle	-0.1165	0.01596
Non-Hodgkin Lymphoma	Smoking Status: Current[5]	Lifestyle	0.3462	9.278e-08
Non-Hodgkin Lymphoma	Happiness with Own Health[21]	Lifestyle	1.33	1.388e-12
Non-Hodgkin Lymphoma	Pack Years of Smoking[5]	Lifestyle	1.184	0.02551
Non-Hodgkin Lymphoma	ICD: Adverse reaction systemic agents[22]	EHR	0.417	0.3438
Non-Hodgkin Lymphoma	ICD10: Other Medical Care[23]	EHR	0.6063	6.108e-05
Non-Hodgkin Lymphoma	OPCS: Other Excision of Breast[24]	EHR	0.7158	5.257e-08
Non-Hodgkin Lymphoma	OPCS: Dissection of lymph nodes[25]	EHR	0.328	0.07675
Non-Hodgkin Lymphoma	OPCS: Intravenous Injection	EHR	-0.02045	0.9093
Kidney Cancer	Cyastin C[9]	Biomarker	2.371	0.07692
Kidney Cancer	HDL Cholesterol[17]	Biomarker	-2.604	0.0003064
Kidney Cancer	Urate[26]	Biomarker	0.3856	0.4879
Kidney Cancer	Impedance of Whole Body[27]	Medical	-0.0219	0.9786
Kidney Cancer	Sex[28]	Medical	0.3571	0.009129
Kidney Cancer	Age[3]	Lifestyle	-1.964	1.588e-13
Kidney Cancer	Waist Circumference[29]	Medical	2.129	0.01338
Kidney Cancer	Number of treatments/medications taken[4]	Medical	2.004	0.009762
Kidney Cancer	Pack Years of Smoking[5]	Lifestyle	2.008	0.01199
Kidney Cancer	Medication: Amlodipine[30]	Medical	0.1999	0.2344
Kidney Cancer	OPCS: Endoscopic Examination of Bladder[31]	EHR	0.3422	0.03669
Pancreatic Cancer	Workplace Very Noisy[32]	Lifestyle	-0.3584	0.03749

Pancreatic Cancer	Cyastin C[9]	Biomarker	-0.2551	0.907
Pancreatic Cancer	HbA1c[33]	Biomarker	6.159	0.0006717
Pancreatic Cancer	Age[3]	Lifestyle	-3.113	1.417e-22
Pancreatic Cancer	Smoking Status: Current[5]	Lifestyle	0.2634	0.1146
Pancreatic Cancer	Pack Years of Smoking[5]	Lifestyle	2.852	0.001477
Pancreatic Cancer	ICD: Problems Related to Lifestyle[6]	EHR	-0.07094	0.8231
Pancreatic Cancer	PGS: Prostate[13]	PGS	3.28	1.141e-12
Pancreatic Cancer	OPCS: Knee replacement[34]	EHR	0.2199	0.628
Colon Cancer	Illness of Father: Stroke[35]	Medical	0.3246	0.004137
Colon Cancer	Sex[36]	Medical	0.01862	0.8378
Colon Cancer	Age[3]	Lifestyle	-2.918	6.94e-64
Colon Cancer	Waist Circumference[37]	Medical	1.946	1.688e-05
Colon Cancer	Standing Height[38]	Medical	1.237	0.04749
Colon Cancer	Pulse Rate[39]	Medical	1.229	0.0004841
Colon Cancer	Smoking Status: Never[5]	Lifestyle	-0.1131	0.0806
Colon Cancer	Happiness with Own Health[21]	Lifestyle	1.02	0.0004715
Colon Cancer	Pack Years of Smoking[5]	Lifestyle	-0.3637	0.6528
Colon Cancer	PGS: Oral/Pharynx[13]	PGS	2.148	8.271e-17
Colon Cancer	Medication: Metformin[40]	Medical	0.0269	0.8558
Melanoma	Vitamin D[41]	Biomarker	3.03	7.248e-36
Melanoma	Sex[42]	Medical	0.1979	8.122e-06
Melanoma	Age[3]	Lifestyle	-2.882	9.576e-227
Melanoma	Standing Height[43]	Medical	1.996	5.496e-10
Melanoma	Ever had PSA Test[44]	Medical	0.1147	0.007904
Melanoma	Skin Color[45]	Medical	-1.082	1.054e-13
Melanoma	Census: White Population[45]	Census	0.2566	0.324
Melanoma	Census: Highest Education Apprenticeship[46]	Census	0.5513	6.043e-06
Melanoma	Census: Provides 1-19 Hours Unpaid Care	Census	0.5169	0.001059
Melanoma	ICD: Transplanted Organ[47]	EHR	1.833	4.433e-16
Melanoma	PGS: Pancreas[13]	PGS	1.485	4.253e-26
Prostate Cancer	Illness of Father: Prostate Cancer[48]	Medical	0.477	1.175e-11
Prostate Cancer	Vitamin D[49]	Biomarker	1.084	0.001333
Prostate Cancer	IGF-1[50]	Biomarker	1.89	4.472e-05
Prostate Cancer	Own Outright Accommodation Lived In[51]	Lifestyle	0.012	0.8039
Prostate Cancer	Age[3]	Lifestyle	-2.933	1.182e-105
Prostate Cancer	Ever had Bowel Cancer Screening[52]	Medical	0.05823	0.188
Prostate Cancer	Ever had PSA Test[52]	Medical	0.3188	1.402e-12
Prostate Cancer	Alcohol Usually Taken with Meal[53]	Lifestyle	0.09124	0.09644
Prostate Cancer	Census: Self-employed Individuals[54]	Census	0.7887	2.711e-06
Prostate Cancer	PGS: Leukemia[13]	PGS	3.182	2.315e-57
Breast Cancer	Illness of Siblings: Breast Cancer[55]	Medical	0.5443	1.802e-17
Breast Cancer	Testosterone[56]	Biomarker	7.635	2.372e-09
Breast Cancer	SHBG[56]	Biomarker	-0.5132	0.005782
Breast Cancer	Age[3]	Lifestyle	-0.8004	5.869e-10
Breast Cancer	Standing Height[57]	Medical	3.104	2.184e-11
Breast Cancer	Comparative Body Size Age 10[58]	Medical	-0.1574	0.006711
Breast Cancer	Ever had Breast Cancer Screening[59]	Medical	0.08622	0.2532
Breast Cancer	Happiness with Own Health[21]	Lifestyle	1.126	9.582e-17
Breast Cancer	PGS: Endometrial[13]	PGS	2.317	1.343e-25
Breast Cancer	Medication: Arimidex[60]	Medical	1.504	0.002682
Lymphocytic Leukemia	White Blood Cell[61]	Biomarker	46.23	1.635e-33
Lymphocytic Leukemia	Red Blood Cell Count[61]	Biomarker	1.873	0.1912
Lymphocytic Leukemia	Platelet Count[61]	Biomarker	-10.65	5.381e-06
Lymphocytic Leukemia	Mean corpuscular hemoglobin	Biomarker	-2.66	0.07905

Lymphocytic Leukemia	C-reactive protein[62]	Biomarker	-0.07948	0.9474
Lymphocytic Leukemia	Whole Body Fat Mass[63]	Medical	-1.067	0.5375
Lymphocytic Leukemia	BMI by Impedance[64]	Medical	2.649	0.1492
Lymphocytic Leukemia	Age[3]	Lifestyle	-3.003	5.159e-14
Lymphocytic Leukemia	Pulse Rate[39]	Medical	-0.353	0.6808
Lymphocytic Leukemia	Smokers in Household: More Than One[65]	Lifestyle	0.2009	0.4285
Lymphocytic Leukemia	Smoking Status: Current[5]	Lifestyle	0.1284	0.6581
Lymphocytic Leukemia	OPCS: Diagnostic Puncture of Bone[61]	EHR	-0.7424	0.4292
Stomach	Cyastin C[9]	Biomarker	-0.6365	0.7272
Stomach	HbA1c[33]	Biomarker	5.343	0.03658
Stomach	Average Household Income[66]	Lifestyle	-0.2772	0.139
Stomach	Sex[67]	Medical	0.708	7.463e-11
Stomach	Age[3]	Lifestyle	-2.544	2.997e-19
Stomach	Waist Circumference[68]	Medical	1.878	0.0064
Stomach	Alcohol Usually Taken with Meal[69]	Lifestyle	-0.03657	0.7856
Stomach	Age[3] at First Birth	Medical	-1.327	0.04835
Stomach	Smoking Status: Never[5]	Lifestyle	-0.4367	1.892e-05
Stomach	Pack Years of Smoking[5]	Lifestyle	3.021	1.855e-05
Stomach	OPCS: Endoscopic Extirpation of Bladder	EHR	0.6067	0.1469
All Leukemia	Platelet Count[70]	Biomarker	1.442	0.3463
All Leukemia	Lymphocyte Count[70]	Biomarker	98.84	1.999e-66
All Leukemia	Eosinophil Count[70]	Biomarker	-9.429	0.02352
All Leukemia	Calcium[71]	Biomarker	-1.886	0.1837
All Leukemia	Cholesterol[17]	Biomarker	-3.328	5.134e-07
All Leukemia	Sex[72]	Medical	0.1185	0.4421
All Leukemia	Age[3]	Lifestyle	-3.362	6.904e-30
All Leukemia	Standing Height[73]	Medical	4.379	4.31e-05
All Leukemia	OPCS: Diagnostic Puncture of Bone[74]	EHR	2.531	5.321e-22
All Leukemia	OPCS: Blood Withdrawl	EHR	0.5891	0.1675

Table 11. The features employed in the 10 feature linear models. The 10 features were determined by an elastic net logistic regression, and then the coefficients were compute within a traditional logistic regression with only the 10 features made available to the model. The P-values are computed from Wald tests.

Cancer	Feature	Type	Importance
Lung Cancer	Workplace Very Noisy	Lifestyle	0.006854
Lung Cancer	Job Involved Shift Work	Lifestyle	0.006329
Lung Cancer	Age	Lifestyle	0.006567
Lung Cancer	Ever had Breast Cancer Screening	Medical	0.005592
Lung Cancer	Smoking Status: Never	Lifestyle	0.00729
Lung Cancer	Smoking Status: Previous	Lifestyle	0.006863
Lung Cancer	Smoking Status: Current	Lifestyle	0.01361
Lung Cancer	Alcohol Intake: 1-3x Month	Lifestyle	0.008229
Lung Cancer	Alcohol Intake: Special Occasions	Lifestyle	0.007099
Lung Cancer	Pack Years of Smoking	Lifestyle	0.03167
Endometrial Cancer	Cyastin C	Biomarker	0.01784
Endometrial Cancer	IGF-1	Biomarker	0.01223
Endometrial Cancer	SHBG	Biomarker	0.01487
Endometrial Cancer	BMI by Impedance	Medical	0.01541
Endometrial Cancer	Waist Circumfrence	Medical	0.01236
Endometrial Cancer	Had Menopause	Medical	0.01203
Endometrial Cancer	Smoking Status: Never	Lifestyle	0.01164
Endometrial Cancer	Pack Years of Smoking	Lifestyle	0.0123
Endometrial Cancer	Census: Asian Individuals	Census	0.01231
Endometrial Cancer	Census: Highest Education Apprenticeship	Census	0.01326
Bladder Cancer	Weight by Impedance	Medical	0.03093
Bladder Cancer	Impedance of Whole Body	Medical	0.03079
Bladder Cancer	Ever Used Hormone Replacement	Medical	0.1218
Bladder Cancer	Ever had Pap Smear	Medical	0.1295
Bladder Cancer	Age at First Birth	Medical	0.03161
Bladder Cancer	Smokers in Household: More Than One	Lifestyle	0.04409
Bladder Cancer	Smoking Status: Never	Lifestyle	0.04913
Bladder Cancer	Smoking Status: Current	Lifestyle	0.03033
Bladder Cancer	Census: No Qualifications	Census	0.03421
Bladder Cancer	PGS: Ovary	PGS	0.03766
Non-Hodgkin Lymphoma	Workplace full of chemical fumes	Lifestyle	0.02928
Non-Hodgkin Lymphoma	Job Involved Shift Work	Lifestyle	0.02354
Non-Hodgkin Lymphoma	Leisure Activity: Pub or Social Club	Lifestyle	0.02434
Non-Hodgkin Lymphoma	Age	Lifestyle	0.06431
Non-Hodgkin Lymphoma	Smoking Status: Never	Lifestyle	0.03948
Non-Hodgkin Lymphoma	Smoking Status: Current	Lifestyle	0.03187
Non-Hodgkin Lymphoma	General Happiness	Lifestyle	0.024
Non-Hodgkin Lymphoma	Happiness with Own Health	Lifestyle	0.03198
Non-Hodgkin Lymphoma	Pack Years of Smoking	Lifestyle	0.06421
Non-Hodgkin Lymphoma	Belief that own life is meaningful	Lifestyle	0.02612
Kidney Cancer	Cyastin C	Biomarker	0.09667
Kidney Cancer	Weight by Impedance	Medical	0.05944
Kidney Cancer	Impedance of Whole Body	Medical	0.0461
Kidney Cancer	Sex	Medical	0.0299
Kidney Cancer	Age	Lifestyle	0.1518
Kidney Cancer	Waist Circumfrence	Medical	0.03911
Kidney Cancer	Number of treatments/medications taken	Medical	0.05081
Kidney Cancer	Ever Used Hormone Replacement	Medical	0.1175
Kidney Cancer	Ever had Pap Smear	Medical	0.0944
Kidney Cancer	Pack Years of Smoking	Lifestyle	0.07084
Pancreatic Cancer	Job Involved Shift Work	Lifestyle	0.01711
Pancreatic Cancer	Cyastin C	Biomarker	0
Pancreatic Cancer	HbA1c	Biomarker	0.08127

Pancreatic Cancer	Age	Lifestyle	0.5191
Pancreatic Cancer	Age Started Oral Contraceptive	Medical	0.1205
Pancreatic Cancer	Ever Used Hormone Replacement	Medical	0.1183
Pancreatic Cancer	General Happiness	Lifestyle	0.008237
Pancreatic Cancer	Happiness with Own Health	Lifestyle	0.002338
Pancreatic Cancer	Pack Years of Smoking	Lifestyle	0.1128
Pancreatic Cancer	PGS: Prostate	PGS	0.02038
Colon Cancer	Illness of Father: Stroke	Medical	0.01317
Colon Cancer	Age	Lifestyle	0.7503
Colon Cancer	Waist Circumfrence	Medical	0.01216
Colon Cancer	Pulse Rate	Medical	0.0138
Colon Cancer	Age Started Oral Contraceptive	Medical	0.07752
Colon Cancer	Smoking Status: Never	Lifestyle	0.004363
Colon Cancer	Happiness with Own Health	Lifestyle	0.01822
Colon Cancer	Pack Years of Smoking	Lifestyle	0.02082
Colon Cancer	Census: Other Ethnicity Individuals	Census	0.01895
Colon Cancer	PGS: Oral/Pharynx	PGS	0.04238
Melanoma	Vitamin D	Biomarker	0.0382
Melanoma	Testosterone	Biomarker	0.03677
Melanoma	Age	Lifestyle	0.08613
Melanoma	Ever had PSA Test	Medical	0.0383
Melanoma	Skin Color	Medical	0.03306
Melanoma	Use of Sun/UV Protection	Lifestyle	0.02638
Melanoma	Ever Used Hormone Replacement	Medical	0.06083
Melanoma	Census: White Population	Census	0.0335
Melanoma	Census: Highest Education Apprenticeship	Census	0.05053
Melanoma	PGS: Pancreas	PGS	0.02996
Prostate Cancer	Vitamin D	Biomarker	0.01502
Prostate Cancer	Testosterone	Biomarker	0.01461
Prostate Cancer	IGF-1	Biomarker	0.01725
Prostate Cancer	Live in Apartment	Lifestyle	0.0149
Prostate Cancer	Own Outright Accommodation Lived In	Lifestyle	0.01692
Prostate Cancer	Number Vehicles in House	Lifestyle	0.01895
Prostate Cancer	Age	Lifestyle	0.07998
Prostate Cancer	Ever had PSA Test	Medical	0.01708
Prostate Cancer	Belief that own life is meaningful	Lifestyle	0.01502
Prostate Cancer	PGS: Leukemia	PGS	0.01747
Breast Cancer	Illness of Siblings: Breast Cancer	Medical	0.02766
Breast Cancer	Testosterone	Biomarker	0.01363
Breast Cancer	SHBG	Biomarker	0.006242
Breast Cancer	Age	Lifestyle	0.01935
Breast Cancer	Hip Circumfrence	Medical	0.03379
Breast Cancer	Comparative Body Size Age 10	Medical	0.0107
Breast Cancer	Ever had Breast Cancer Screening	Medical	0.03667
Breast Cancer	Happiness with Own Health	Lifestyle	0.08438
Breast Cancer	Pack Years of Smoking	Lifestyle	0.1986
Breast Cancer	PGS: Endometrial	PGS	0.525
Lymphocytic Leukemia	White Blood Cell	Biomarker	0.1624
Lymphocytic Leukemia	Red Blood Cell Count	Biomarker	0.01262
Lymphocytic Leukemia	Platelet Count	Biomarker	0.03289
Lymphocytic Leukemia	Vitamin D	Biomarker	0.01155
Lymphocytic Leukemia	Mean corpuscular hemoglobin	Biomarker	0.01599
Lymphocytic Leukemia	C-reactive protein	Biomarker	0.01843
Lymphocytic Leukemia	Age	Lifestyle	0.01439
Lymphocytic Leukemia	Age Started Oral Contraceptive	Medical	0.0115

Lymphocytic Leukemia	PGS: Kidney Cancer	PGS	0.01155
Lymphocytic Leukemia	PGS: Ovary	PGS	0.01674
Stomach	Job Involved Shift Work	Lifestyle	0.007741
Stomach	Number Vehicles in House	Lifestyle	0.005791
Stomach	Age	Lifestyle	0.008559
Stomach	Getting Up in Morning	Lifestyle	0.006975
Stomach	Ever Used Hormone Replacement	Medical	0.006585
Stomach	Alchol Intake Compared 10 Years Prior	Lifestyle	0.00626
Stomach	Ever had Pap Smear	Medical	0.01034
Stomach	Smoking Status: Never	Lifestyle	0.006532
Stomach	Census: Individuals 10 to 14 Years	Census	0.006173
Stomach	OPCS: Excision of Lesion of Skin	EHR	0.005951
All Leukemia	Lymphocyte Count	Biomarker	0.02378
All Leukemia	Eosinophil Count	Biomarker	0.007597
All Leukemia	Cyastin C	Biomarker	0.008005
All Leukemia	HDL Cholesterol	Biomarker	0.007698
All Leukemia	Age	Lifestyle	0.007635
All Leukemia	Water Intake	Diet	0.008799
All Leukemia	Census: Aged 8-9 Individuals	Census	0.007679
All Leukemia	Census: Renting Individuals	Census	0.0101
All Leukemia	Census: Daily Activities Limited a Lot	Census	0.01049
All Leukemia	PGS: Prostate Cancer	PGS	0.007865

Table 12. The machine learning features that overalap the 10-feature linear model features. The importances are extracted from the “feature_importances_” attribute which is computed when each machine learning model is fit with the sklearn library.

Feature	Cancer	Linear Weight	ML Weight
Workplace Very Noisy	Lung Cancer	0.04329	0.06847
Job Involved Shift Work	Lung Cancer	0.005591	0.06322
Age	Lung Cancer	0.2216	0.06561
Ever had Breast Cancer Screening	Lung Cancer	0.02888	0.05586
Smoking Status: Never	Lung Cancer	0.1267	0.07283
Smoking Status: Previous	Lung Cancer	0.06312	0.06856
Smoking Status: Current	Lung Cancer	0.01886	0.136
Alcohol Intake: 1-3x Month	Lung Cancer	0.009937	0.08221
Alcohol Intake: Special Occasions	Lung Cancer	0.01	0.07093
Pack Years of Smoking	Lung Cancer	0.472	0.3164
Cystatin C	Endometrial Cancer	0.01547	0.1329
IGF-1	Endometrial Cancer	0.141	0.09113
SHBG	Endometrial Cancer	0.1788	0.1108
BMI by Impedance	Endometrial Cancer	0.3675	0.1148
Waist Circumfrence	Endometrial Cancer	0.01021	0.09204
Had Menopause	Endometrial Cancer	0.07878	0.0896
Smoking Status: Never	Endometrial Cancer	0.01417	0.08667
Pack Years of Smoking	Endometrial Cancer	0.1047	0.09161
Census: Asian Individuals	Endometrial Cancer	0.03263	0.09169
Census: Highest Education Apprenticeship	Endometrial Cancer	0.05683	0.09878
Weight by Impedance	Bladder Cancer	0.04686	0.05727
Impedance of Whole Body	Bladder Cancer	0.5009	0.05702
Ever Used Hormone Replacement	Bladder Cancer	0.07271	0.2255
Ever had Pap Smear	Bladder Cancer	0.025	0.2398
Age at First Birth	Bladder Cancer	0.19	0.05853
Smokers in Household: More Than One	Bladder Cancer	0.01467	0.08165
Smoking Status: Never	Bladder Cancer	0.07053	0.09097
Smoking Status: Current	Bladder Cancer	0.02785	0.05617
Census: No Qualifications	Bladder Cancer	0.01744	0.06334
PGS: Ovary	Bladder Cancer	0.03402	0.06973
Workplace full of chemical fumes	Non-Hodgkin Lymphoma	0.04852	0.08153
Job Involved Shift Work	Non-Hodgkin Lymphoma	0.02598	0.06556
Leisure Activity: Pub or Social Club	Non-Hodgkin Lymphoma	0.002281	0.06777
Age	Non-Hodgkin Lymphoma	0.2357	0.1791
Smoking Status: Never	Non-Hodgkin Lymphoma	0.01698	0.1099
Smoking Status: Current	Non-Hodgkin Lymphoma	0.05019	0.08875
General Happiness	Non-Hodgkin Lymphoma	0.1496	0.06682
Happiness with Own Health	Non-Hodgkin Lymphoma	0.2438	0.08904
Pack Years of Smoking	Non-Hodgkin Lymphoma	0.1836	0.1788
Belief that own life is meaningful	Non-Hodgkin Lymphoma	0.04344	0.07274
Cystatin C	Kidney Cancer	0.2407	0.1278
Weight by Impedance	Kidney Cancer	0.03397	0.07856
Impedance of Whole Body	Kidney Cancer	0.01718	0.06093
Sex	Kidney Cancer	0.03795	0.03952
Age	Kidney Cancer	0.1343	0.2007
Waist Circumfrence	Kidney Cancer	0.1896	0.05169
Number of treatments/medications taken	Kidney Cancer	0.1816	0.06716
Ever Used Hormone Replacement	Kidney Cancer	0.01202	0.1553
Ever had Pap Smear	Kidney Cancer	0.0006992	0.1248
Pack Years of Smoking	Kidney Cancer	0.152	0.09363
Job Involved Shift Work	Pancreatic Cancer	0.004978	0.01711
Cystatin C	Pancreatic Cancer	0.008667	0
HbA1c	Pancreatic Cancer	0.3894	0.08127

Age	Pancreatic Cancer	0.1863	0.5191
Age Started Oral Contraceptive	Pancreatic Cancer	0.02394	0.1205
Ever Used Hormone Replacement	Pancreatic Cancer	0.01881	0.1183
General Happiness	Pancreatic Cancer	0.01854	0.008237
Happiness with Own Health	Pancreatic Cancer	0.03224	0.002338
Pack Years of Smoking	Pancreatic Cancer	0.1174	0.1128
PGS: Prostate	Pancreatic Cancer	0.1996	0.02038
Illness of Father: Stroke	Colon Cancer	0.03058	0.01355
Age	Colon Cancer	0.253	0.7722
Waist Circumfrence	Colon Cancer	0.2276	0.01252
Pulse Rate	Colon Cancer	0.1012	0.0142
Age Started Oral Contraceptive	Colon Cancer	0.04065	0.07978
Smoking Status: Never	Colon Cancer	0.01129	0.00449
Happiness with Own Health	Colon Cancer	0.07956	0.01875
Pack Years of Smoking	Colon Cancer	0.001506	0.02143
Census: Other Ethnicity Individuals	Colon Cancer	0.05451	0.01951
PGS: Oral/Pharynx	Colon Cancer	0.2	0.04361
Vitamin D	Melanoma	0.2377	0.08809
Testosterone	Melanoma	0.1616	0.08478
Age	Melanoma	0.2326	0.1986
Ever had PSA Test	Melanoma	0.01156	0.08832
Skin Color	Melanoma	0.08176	0.07622
Use of Sun/UV Protection	Melanoma	0.03901	0.06084
Ever Used Hormone Replacement	Melanoma	0.006615	0.1403
Census: White Population	Melanoma	0.0612	0.07726
Census: Highest Education Apprenticeship	Melanoma	0.05147	0.1165
PGS: Pancreas	Melanoma	0.1165	0.06908
Vitamin D	Prostate Cancer	0.1113	0.0661
Testosterone	Prostate Cancer	0.001814	0.06429
IGF-1	Prostate Cancer	0.2011	0.07591
Live in Apartment	Prostate Cancer	0.005465	0.06559
Own Outright Accommodation Lived In	Prostate Cancer	0.001806	0.07449
Number Vehicles in House	Prostate Cancer	0.01508	0.08341
Age	Prostate Cancer	0.2932	0.352
Ever had PSA Test	Prostate Cancer	0.03626	0.07519
Belief that own life is meaningful	Prostate Cancer	0.01273	0.06612
PGS: Leukemia	Prostate Cancer	0.3214	0.07688
Illness of Siblings: Breast Cancer	Breast Cancer	0.03958	0.02894
Testosterone	Breast Cancer	0.5318	0.01426
SHBG	Breast Cancer	0.02173	0.00653
Age	Breast Cancer	0.04849	0.02024
Hip Circumfrence	Breast Cancer	0.06231	0.03535
Comparative Body Size Age 10	Breast Cancer	0.01363	0.01119
Ever had Breast Cancer Screening	Breast Cancer	0.006698	0.03836
Happiness with Own Health	Breast Cancer	0.07894	0.08827
Pack Years of Smoking	Breast Cancer	0.03309	0.2077
PGS: Endometrial	Breast Cancer	0.1637	0.5492
White Blood Cell	Lymphocytic Leukemia	0.6416	0.5272
Red Blood Cell Count	Lymphocytic Leukemia	0.03743	0.04095
Platelet Count	Lymphocytic Leukemia	0.1603	0.1067
Vitamin D	Lymphocytic Leukemia	0.01153	0.0375
Mean corpuscular hemoglobin	Lymphocytic Leukemia	0.03743	0.05191
C-reactive protein	Lymphocytic Leukemia	0.006552	0.05982
Age	Lymphocytic Leukemia	0.04242	0.0467
Age Started Oral Contraceptive	Lymphocytic Leukemia	0.005205	0.03732

PGS: Kidney Cancer	Lymphocytic Leukemia	0.004986	0.0375
PGS: Ovary	Lymphocytic Leukemia	0.05263	0.05435
Job Involved Shift Work	Stomach	0.05366	0.1092
Number Vehicles in House	Stomach	0.08322	0.08167
Age	Stomach	0.3899	0.1207
Getting Up in Morning	Stomach	0.1144	0.09837
Ever Used Hormone Replacement	Stomach	0.01721	0.09287
Alchol Intake Compared 10 Years Prior	Stomach	0.03487	0.08828
Ever had Pap Smear	Stomach	0.02164	0.1458
Smoking Status: Never	Stomach	0.1017	0.09213
Census: Individuals 10 to 14 Years	Stomach	0.1584	0.08706
OPCS: Excision of Lesion of Skin	Stomach	0.025	0.08393
Lymphocyte Count	All Leukemia	0.7615	0.2386
Eosinophil Count	All Leukemia	0.0908	0.07624
Cystatin C	All Leukemia	0.0282	0.08033
HDL Cholesterol	All Leukemia	0.03491	0.07725
Age	All Leukemia	0.02539	0.07662
Water Intake	All Leukemia	0.02044	0.08829
Census: Aged 8-9 Individuals	All Leukemia	0.002844	0.07706
Census: Renting Individuals	All Leukemia	0.01548	0.1014
Census: Daily Activities Limited a Lot	All Leukemia	0.007308	0.1052
PGS: Prostate Cancer	All Leukemia	0.01316	0.07892

Table 13. The feature weights for machine learning and linear models. The weights for the linear models were the absolute value of the linear coefficients (Table 8) divided by the sum of the absolute value of the linear coefficients across all features in a cancer. Similarly, the machine learning weights were the machine learning feature importances divided by the sum of the feature importances across all features in a cancer. The NA values occur when a linear feature was not used in the machine learning model. Only the 8 most common features across the two models are displayed here.

Cancer	Feature - 1	Feature - 2	Coef.	SE	P
Lung Cancer	Smoking Status: Never	Smoking Status: Current	1.08	0.0644	2.85e-63
Lung Cancer	Smoking Status: Previous	Smoking Status: Current	2.45	0.0831	1.62e-191
Lung Cancer	Pack Years of Smoking	General Happiness	-17.5	2.09	4.97e-17
Lung Cancer	Happiness with Own Health	Belief that own life is meaningful	4.89	0.835	4.99e-09
Non-Hodgkin Lymphoma	Testosterone	Age	-6.49	1	1e-10
Lymphocytic Leukemia	White Blood Cell	White Blood Cell	44.1	3.5	2.99e-36
All Leukemia	Lymphocyte Count	Lymphocyte Count	97.3	5.52	1.33e-69
All Leukemia	Lymphocyte Count	Monocyte Count	-355	58.4	1.16e-09
All Leukemia	White Blood Cell	Lymphocyte Count	-246	19.5	3.44e-36

Table 14. The significance of feature interactions. All interactions listed by the xgbfir package applied to the XGBoost model were examined in a linear model which included both univariate terms (of the interaction) and the interaction. The coefficient, standard error and Wald's test P-value for the interaction term in that linear regression is recorded here.

Cancer	Effector	Coefficient	SE	P
Lung Cancer	Time From Attend	-1.54e-06	3.42e-06	0.651
Lung Cancer	Age at Diagnosis	0.00116	0.000626	0.0633
Lung Cancer	Age at Assessment	0.00161	0.000672	0.0169
Lung Cancer	Health Rating	-0.144	0.032	7.12e-06
Lung Cancer	Income	0.115	0.0449	0.0102
Endometrial Cancer	Time From Attend	7.43e-06	7.63e-06	0.331
Endometrial Cancer	Age at Diagnosis	0.00667	0.00117	2.61e-08
Endometrial Cancer	Age at Assessment	0.00746	0.00119	9.53e-10
Endometrial Cancer	Health Rating	0.0186	0.0198	0.348
Endometrial Cancer	Income	0.883	0.124	4.95e-12
Bladder Cancer	Time From Attend	-3.21e-06	3.77e-06	0.395
Bladder Cancer	Age at Diagnosis	0.00975	0.000571	7.05e-55
Bladder Cancer	Age at Assessment	0.013	0.00058	3.24e-85
Bladder Cancer	Health Rating	-0.208	0.0365	1.62e-08
Bladder Cancer	Income	0.215	0.0553	0.000108
Non-Hodgkin Lymphoma	Time From Attend	-2e-06	2.26e-06	0.376
Non-Hodgkin Lymphoma	Age at Diagnosis	0.00905	0.000301	1.94e-168
Non-Hodgkin Lymphoma	Age at Assessment	0.0111	0.000315	5.05e-218
Non-Hodgkin Lymphoma	Health Rating	0.317	0.0613	2.6e-07
Non-Hodgkin Lymphoma	Income	0.00858	0.0065	0.187
Kidney Cancer	Time From Attend	-7.88e-06	6.52e-06	0.227
Kidney Cancer	Age at Diagnosis	0.00844	0.00099	2.74e-16
Kidney Cancer	Age at Assessment	0.0111	0.000998	1.05e-25
Kidney Cancer	Health Rating	-0.277	0.0573	1.91e-06
Kidney Cancer	Income	0.159	0.0792	0.0455
Pancreatic Cancer	Time From Attend	-1.72e-05	8.33e-06	0.0399
Pancreatic Cancer	Age at Diagnosis	0.0163	0.0012	9.15e-34
Pancreatic Cancer	Age at Assessment	0.023	0.00111	1.21e-63
Pancreatic Cancer	Health Rating	-0.285	0.11	0.00991
Pancreatic Cancer	Income	0.236	0.109	0.0308
Colon Cancer	Time From Attend	-8.32e-06	4.1e-06	0.0425
Colon Cancer	Age at Diagnosis	0.0173	0.000435	2.83e-215
Colon Cancer	Age at Assessment	0.0217	0.000347	0
Colon Cancer	Health Rating	-0.0593	0.0197	0.0027
Colon Cancer	Income	0.207	0.0597	0.000554
Melanoma	Time From Attend	-3.7e-06	1.75e-06	0.0347
Melanoma	Age at Diagnosis	0.0154	0.000163	0
Melanoma	Age at Assessment	0.0185	0.000127	0
Melanoma	Health Rating	-0.0109	0.00525	0.038
Melanoma	Income	0.323	0.0262	2.18e-34
Prostate Cancer	Time From Attend	-1.92e-05	2.54e-06	6.81e-14
Prostate Cancer	Age at Diagnosis	0.0146	0.000394	5.09e-238
Prostate Cancer	Age at Assessment	0.0192	0.000342	0
Prostate Cancer	Health Rating	-5.28	1.38	0.000126
Prostate Cancer	Income	-0.0727	0.0148	9.74e-07
Breast Cancer	Time From Attend	-2.37e-06	1.75e-06	0.175
Breast Cancer	Age at Diagnosis	0.00424	0.00024	2.59e-65
Breast Cancer	Age at Assessment	0.00494	0.000242	2.55e-85
Breast Cancer	Health Rating	0.0101	0.00469	0.031
Breast Cancer	Income	0.202	0.0288	3.12e-12
Lymphocytic Leukemia	Time From Attend	-1.13e-05	5.95e-06	0.059
Lymphocytic Leukemia	Age at Diagnosis	-0.00175	0.00103	0.0918
Lymphocytic Leukemia	Age at Assessment	-0.000676	0.00102	0.51

Lymphocytic Leukemia	Health Rating	-0.0306	0.0551	0.579
Lymphocytic Leukemia	Income	-0.0623	0.0774	0.422
Stomach	Time From Attend	-1.37e-05	5.34e-06	0.0107
Stomach	Age at Diagnosis	0.00891	0.000925	3.57e-20
Stomach	Age at Assessment	0.0118	0.000884	2.58e-35
Stomach	Health Rating	-0.0262	0.0272	0.336
Stomach	Income	0.367	0.0748	1.21e-06
All Leukemia	Time From Attend	-2.18e-05	4.07e-06	1.4e-07
All Leukemia	Age at Diagnosis	-0.00177	0.000706	0.0124
All Leukemia	Age at Assessment	-9.25e-05	0.000828	0.911
All Leukemia	Health Rating	0.167	0.117	0.155
All Leukemia	Income	-0.00547	0.0119	0.645

Table 15. The summary statistics from associations between the effectors and the linear predictors

Cancer	Effector	Coefficient	SE	P
Lung Cancer	Time From Attend	6.77e-06	2.92e-06	0.0207
Lung Cancer	Age at Diagnosis	-0.00249	0.000533	3.34e-06
Lung Cancer	Age at Assessment	-0.00398	0.000575	7.58e-12
Lung Cancer	Health Rating	0.118	0.0279	2.32e-05
Lung Cancer	Income	-0.0188	0.0392	0.632
Endometrial Cancer	Time From Attend	-3.14e-07	4.3e-06	0.942
Endometrial Cancer	Age at Diagnosis	-0.002	0.00068	0.00348
Endometrial Cancer	Age at Assessment	-0.00246	0.000684	0.000353
Endometrial Cancer	Health Rating	-0.00568	0.011	0.607
Endometrial Cancer	Income	-0.325	0.0715	7.28e-06
Bladder Cancer	Time From Attend	2.74e-06	4.33e-06	0.526
Bladder Cancer	Age at Diagnosis	-0.00944	0.000693	1.13e-37
Bladder Cancer	Age at Assessment	-0.0124	0.000718	2e-56
Bladder Cancer	Health Rating	0.266	0.0411	1.7e-10
Bladder Cancer	Income	-0.183	0.0629	0.00367
Non-Hodgkin Lymphoma	Time From Attend	4.81e-06	2.34e-06	0.0396
Non-Hodgkin Lymphoma	Age at Diagnosis	-0.00703	0.000336	1.45e-89
Non-Hodgkin Lymphoma	Age at Assessment	-0.00886	0.000358	7.9e-121
Non-Hodgkin Lymphoma	Health Rating	-0.323	0.0633	3.73e-07
Non-Hodgkin Lymphoma	Income	-0.00908	0.00672	0.177
Kidney Cancer	Time From Attend	8.08e-06	9.49e-06	0.395
Kidney Cancer	Age at Diagnosis	-0.0136	0.00141	7.39e-20
Kidney Cancer	Age at Assessment	-0.0175	0.00143	1.03e-29
Kidney Cancer	Health Rating	0.546	0.0823	9.51e-11
Kidney Cancer	Income	-0.282	0.116	0.0155
Pancreatic Cancer	Time From Attend	3.02e-05	9.62e-06	0.00183
Pancreatic Cancer	Age at Diagnosis	-0.0199	0.00135	4.17e-38
Pancreatic Cancer	Age at Assessment	-0.0288	0.00112	7.55e-84
Pancreatic Cancer	Health Rating	0.277	0.126	0.0289
Pancreatic Cancer	Income	-0.339	0.124	0.00661
Colon Cancer	Time From Attend	1.05e-05	4.59e-06	0.0227
Colon Cancer	Age at Diagnosis	-0.0182	0.000527	4.24e-178
Colon Cancer	Age at Assessment	-0.0229	0.000448	2.49e-301
Colon Cancer	Health Rating	0.0773	0.022	0.000456
Colon Cancer	Income	-0.241	0.0665	0.000299
Melanoma	Time From Attend	1.1e-06	9e-07	0.22
Melanoma	Age at Diagnosis	-0.00543	0.000121	0
Melanoma	Age at Assessment	-0.00644	0.000121	0
Melanoma	Health Rating	0.00555	0.00268	0.0386
Melanoma	Income	-0.142	0.0134	5.88e-26
Prostate Cancer	Time From Attend	9.94e-06	1.41e-06	2.45e-12
Prostate Cancer	Age at Diagnosis	-0.00493	0.000256	8.49e-77
Prostate Cancer	Age at Assessment	-0.00686	0.000242	1.22e-153
Prostate Cancer	Health Rating	2.06	0.746	0.00573
Prostate Cancer	Income	0.0288	0.00804	0.000348
Breast Cancer	Time From Attend	6.23e-06	2.73e-06	0.0224
Breast Cancer	Age at Diagnosis	-0.00713	0.000371	3.39e-76
Breast Cancer	Age at Assessment	-0.00847	0.000374	5.98e-103
Breast Cancer	Health Rating	-0.0172	0.00736	0.0194
Breast Cancer	Income	-0.398	0.0451	1.81e-18
Lymphocytic Leukemia	Time From Attend	4.27e-05	1.55e-05	0.00641
Lymphocytic Leukemia	Age at Diagnosis	0.000684	0.00273	0.802
Lymphocytic Leukemia	Age at Assessment	-0.00356	0.0027	0.189

Lymphocytic Leukemia	Health Rating	0.0299	0.146	0.838
Lymphocytic Leukemia	Income	0.0816	0.205	0.691
Stomach	Time From Attend	1.57e-05	4.2e-06	0.000203
Stomach	Age at Diagnosis	-0.0058	0.000756	9.94e-14
Stomach	Age at Assessment	-0.00823	0.000746	1.44e-25
Stomach	Health Rating	0.0558	0.0219	0.0111
Stomach	Income	-0.253	0.0608	3.68e-05
All Leukemia	Time From Attend	5.34e-05	9.67e-06	5.68e-08
All Leukemia	Age at Diagnosis	0.00343	0.00168	0.0421
All Leukemia	Age at Assessment	-0.00101	0.00197	0.608
All Leukemia	Health Rating	-0.133	0.279	0.634
All Leukemia	Income	0.00274	0.0283	0.923

Table 16. The summary statistics from associations between the effectors and the linear predictors

References

1. Sprague, B. L. *et al.* Physical activity, white blood cell count, and lung cancer risk in a prospective cohort study. *Cancer Epidemiol. Biomarkers & Prev.* **17**, 2714–2722, DOI: 10.1158/1055-9965.epi-08-0042 (2008).
2. Mao, Y., Hu, J., Ugnat, A.-M., Semenciw, R. & Fincham, S. Socioeconomic status and lung cancer risk in canada. *Int. J. Epidemiol.* **30**, 809–817, DOI: 10.1093/ije/30.4.809 (2001).
3. DePinho, R. A. The age of cancer. *Nature* **408**, 248–254, DOI: 10.1038/35041694 (2000).
4. Onitilo, A. A. *et al.* Diabetes and cancer II: role of diabetes medications and influence of shared risk factors. *Cancer Causes & Control.* **23**, 991–1008, DOI: 10.1007/s10552-012-9971-4 (2012).
5. Jee, S. H., Park, J. Y., Kim, H.-S., Lee, T. Y. & Samet, J. M. White blood cell count and risk for all-cause, cardiovascular, and cancer mortality in a cohort of koreans. *Am. J. Epidemiol.* **162**, 1062–1069, DOI: 10.1093/aje/kwi326 (2005).
6. Kolonel, L. N., Altshuler, D. & Henderson, B. E. The multiethnic cohort study: exploring genes, lifestyle and cancer risk. *Nat. Rev. Cancer* **4**, 519–527, DOI: 10.1038/nrc1389 (2004).
7. KOBAYASHI, S., SUZUKI, S., NIIKAWA, H., SUGAWARA, T. & YANAI, M. Preoperative use of inhaled tiotropium in lung cancer patients with untreated COPD. *Respirology* **14**, 675–679, DOI: 10.1111/j.1440-1843.2009.01543.x (2009).
8. Riise, H. K. R., Riise, T., Natvig, G. K. & Daltveit, A. K. Poor self-rated health associated with an increased risk of subsequent development of lung cancer. *Qual. Life Res.* **23**, 145–153, DOI: 10.1007/s11136-013-0453-2 (2013).
9. Lees, J. S. *et al.* Kidney function and cancer risk: An analysis using creatinine and cystatin c in a cohort study. *EClinicalMedicine* **38**, 101030, DOI: 10.1016/j.eclinm.2021.101030 (2021).
10. Park, S. L. *et al.* Body size, adult BMI gain and endometrial cancer risk: the multiethnic cohort. *Int. J. Cancer* **126**, 490–499, DOI: 10.1002/ijc.24718 (2010).
11. Brinton, L. A. & Felix, A. S. Menopausal hormone therapy and risk of endometrial cancer. *The J. Steroid Biochem. Mol. Biol.* **142**, 83–89, DOI: 10.1016/j.jsbmb.2013.05.001 (2014).
12. Weiss, N. S. & Sayvetz, T. A. Incidence of endometrial cancer in relation to the use of oral contraceptives. *New Engl. J. Medicine* **302**, 551–554, DOI: 10.1056/nejm198003063021004 (1980).
13. Graff, R. E. *et al.* Cross-cancer evaluation of polygenic risk scores for 17 cancer types in two large cohorts. DOI: 10.1101/2020.01.18.911578 (2020).
14. Bernstein, L. *et al.* Tamoxifen therapy for breast cancer and endometrial cancer risk. *JNCI J. Natl. Cancer Inst.* **91**, 1654–1662, DOI: 10.1093/jnci/91.19.1654 (1999).
15. Biggar, R. J., Wohlfahrt, J. & Melbye, M. Digoxin use and the risk of cancers of the corpus uteri, ovary and cervix. *Int. J. Cancer* **131**, 716–721, DOI: 10.1002/ijc.26424 (2011).
16. Doll, K. M. & Winn, A. N. Assessing endometrial cancer risk among US women: long-term trends using hysterectomy-adjusted analysis. *Am. J. Obstet. Gynecol.* **221**, 318.e1–318.e9, DOI: 10.1016/j.ajog.2019.05.024 (2019).
17. Hu, J. *et al.* Dietary cholesterol intake and cancer. *Annals Oncol.* **23**, 491–500, DOI: 10.1093/annonc/mdr155 (2012).
18. ZHANG, Y. Understanding the gender disparity in bladder cancer risk: The impact of sex hormones and liver on bladder susceptibility to carcinogens. *J. Environ. Sci. Heal. Part C* **31**, 287–304, DOI: 10.1080/10590501.2013.844755 (2013).
19. McGrath, M., Michaud, D. S. & Vivo, I. D. Hormonal and reproductive factors and the risk of bladder cancer in women. *Am. J. Epidemiol.* **163**, 236–244, DOI: 10.1093/aje/kwj028 (2005).
20. Chen, C.-L. *et al.* Identification of potential bladder cancer markers in urine by abundant-protein depletion coupled with quantitative proteomics. *J. Proteomics* **85**, 28–43, DOI: 10.1016/j.jprot.2013.04.024 (2013).
21. Liu, B. *et al.* Does happiness itself directly affect mortality? the prospective UK million women study. *The Lancet* **387**, 874–881, DOI: 10.1016/s0140-6736(15)01087-9 (2016).
22. Kobayashi, D. *et al.* Risk factors for adverse reactions from contrast agents for computed tomography. *BMC Med. Informatics Decis. Mak.* **13**, DOI: 10.1186/1472-6947-13-18 (2013).
23. Breslow, R. A., Sorkin, J. D., Frey, C. M. & Kessler, L. G. Americans' knowledge of cancer risk and survival. *Prev. Medicine* **26**, 170–177, DOI: 10.1006/pmed.1996.0136 (1997).
24. Wong, W. W., Schild, S. E., Halyard, M. Y. & Schomberg, P. J. Primary non-hodgkin lymphoma of the breast: The mayo clinic experience. *J. Surg. Oncol.* **80**, 19–25, DOI: 10.1002/jso.10084 (2002).

25. Ha, C. S. *et al.* Localized non-hodgkin lymphoma involving the thyroid gland. *Cancer* **91**, 629–635, DOI: 10.1002/1097-0142(20010215)91:4<629::aid-cncr1045>3.0.co;2-q (2001).

26. Kobylecki, C. J., Afzal, S. & Nordestgaard, B. G. Plasma urate, cancer incidence, and all-cause mortality: A mendelian randomization study. *Clin. Chem.* **63**, 1151–1160, DOI: 10.1373/clinchem.2016.268185 (2017).

27. Thibault, R. & Pichard, C. The evaluation of body composition: A useful tool for clinical practice. *Annals Nutr. Metab.* **60**, 6–16, DOI: 10.1159/000334879 (2012).

28. Scelo, G., Li, P., Chanudet, E. & Muller, D. C. Variability of sex disparities in cancer incidence over 30 years: The striking case of kidney cancer. *Eur. Urol. Focus.* **4**, 586–590, DOI: 10.1016/j.euf.2017.01.006 (2018).

29. Wilson, K. M. & Cho, E. Obesity and kidney cancer. In *Obesity and Cancer*, 81–93, DOI: 10.1007/978-3-319-42542-9_5 (Springer International Publishing, 2016).

30. Rosenberg, L. Calcium channel blockers and the risk of cancer. *JAMA* **279**, 1000, DOI: 10.1001/jama.279.13.1000 (1998).

31. Whittemore, A. S., Paffenbarger, R. S., Anderson, K. & Lee, J. E. Early precursors of urogenital cancers in former college men. *J. Urol.* **132**, 1256–1261, DOI: 10.1016/s0022-5347(17)50118-4 (1984).

32. Forsell, K. *et al.* Work environment and safety climate in the swedish merchant fleet. *Int. Arch. Occup. Environ. Heal.* **90**, 161–168, DOI: 10.1007/s00420-016-1180-0 (2016).

33. Wolpin, B. M. *et al.* Hyperglycemia, insulin resistance, impaired pancreatic -cell function, and risk of pancreatic cancer. *JNCI: J. Natl. Cancer Inst.* **105**, 1027–1035, DOI: 10.1093/jnci/djt123 (2013).

34. Goldacre, M. J., Wotton, C. J., Seagroatt, V. & Yeates, D. Cancer following hip and knee arthroplasty: record linkage study. *Br. J. Cancer* **92**, 1298–1301, DOI: 10.1038/sj.bjc.6602511 (2005).

35. Kyu, H. H. *et al.* Physical activity and risk of breast cancer, colon cancer, diabetes, ischemic heart disease, and ischemic stroke events: systematic review and dose-response meta-analysis for the global burden of disease study 2013. *BMJ* i3857, DOI: 10.1136/bmj.i3857 (2016).

36. Cai, Y. *et al.* Sex differences in colon cancer metabolism reveal a novel subphenotype. *Sci. Reports* **10**, DOI: 10.1038/s41598-020-61851-0 (2020).

37. Moore, L. L. *et al.* BMI and waist circumference as predictors of lifetime colon cancer risk in framingham study adults. *Int. J. Obes.* **28**, 559–567, DOI: 10.1038/sj.ijo.0802606 (2004).

38. Caan, B. *et al.* Body size and the risk of colon cancer in a large case-control study. *Int. J. Obes.* **22**, 178–184, DOI: 10.1038/sj.ijo.0800561 (1998).

39. PERSKLY, V. *et al.* HEART RATE: A RISK FACTOR FOR CANCER? *Am. J. Epidemiol.* **114**, 477–487, DOI: 10.1093/oxfordjournals.aje.a113213 (1981).

40. Nangia-Makker, P. *et al.* Metformin: A potential therapeutic agent for recurrent colon cancer. *PLoS ONE* **9**, e84369, DOI: 10.1371/journal.pone.0084369 (2014).

41. Field, S. & Newton-Bishop, J. A. Melanoma and vitamin d. *Mol. Oncol.* **5**, 197–214, DOI: 10.1016/j.molonc.2011.01.007 (2011).

42. Nosrati, A. & Wei, M. L. Sex disparities in melanoma outcomes: The role of biology. *Arch. Biochem. Biophys.* **563**, 42–50, DOI: 10.1016/j.abb.2014.06.018 (2014).

43. Vena, G. A., Cassano, N., Caccavale, S. & Argenziano, G. Association between melanoma risk and height: A narrative review. *Dermatol. Pract. & Concept.* 82–89, DOI: 10.5826/dpc.0902a02 (2019).

44. Noe, M., Schroy, P., Demierre, M.-F., Babayan, R. & Geller, A. C. Increased cancer risk for individuals with a family history of prostate cancer, colorectal cancer, and melanoma and their associated screening recommendations and practices. *Cancer Causes & Control.* **19**, 1–12, DOI: 10.1007/s10552-007-9064-y (2007).

45. Cust, A. E., Mishra, K. & Berwick, M. Melanoma – role of the environment and genetics. *Photochem. & Photobiol. Sci.* **17**, 1853–1860, DOI: 10.1039/c7pp00411g (2018).

46. Geller, A. C., Swetter, S. M., Brooks, K., Demierre, M.-F. & Yaroch, A. L. Screening, early detection, and trends for melanoma: Current status (2000-2006) and future directions. *J. Am. Acad. Dermatol.* **57**, 555–572, DOI: 10.1016/j.jaad.2007.06.032 (2007).

47. Zwald, F. O. *et al.* Melanoma in solid organ transplant recipients. *Am. J. Transplantation* **10**, 1297–1304, DOI: 10.1111/j.1600-6143.2010.03078.x (2010).

48. Steinberg, G. D., Carter, B. S., Beaty, T. H., Childs, B. & Walsh, P. C. Family history and the risk of prostate cancer. *The Prostate* **17**, 337–347, DOI: 10.1002/pros.2990170409 (1990).

49. CHEN, T. & HOLICK, M. Vitamin d and prostate cancer prevention and treatment. *Trends Endocrinol. Metab.* **14**, 423–430, DOI: 10.1016/j.tem.2003.09.004 (2003).

50. Papatsoris, A. G., Karamouzis, M. V. & Papavassiliou, A. G. Novel insights into the implication of the IGF-1 network in prostate cancer. *Trends Mol. Medicine* **11**, 52–55, DOI: 10.1016/j.molmed.2004.12.005 (2005).

51. Klein, J. & von dem Knesebeck, O. Socioeconomic inequalities in prostate cancer survival: A review of the evidence and explanatory factors. *Soc. Sci. & Medicine* **142**, 9–18, DOI: 10.1016/j.socscimed.2015.07.006 (2015).

52. Hoffman, R. M. Screening for prostate cancer. *New Engl. J. Medicine* **365**, 2013–2019, DOI: 10.1056/nejmcp1103642 (2011).

53. Schoonen, W. M., Salinas, C. A., Kiemeney, L. A. & Stanford, J. L. Alcohol consumption and risk of prostate cancer in middle-aged men. *Int. J. Cancer* **113**, 133–140, DOI: 10.1002/ijc.20528 (2005).

54. Bradley, C. J., Neumark, D., Luo, Z. & Schenk, M. Employment and cancer: Findings from a longitudinal study of breast and prostate cancer survivors. *Cancer Investig.* **25**, 47–54, DOI: 10.1080/07357900601130664 (2007).

55. Laraja, R. D. *et al.* Carcinoma of the breast in three siblings. *Cancer* **55**, 2709–2711, DOI: 10.1002/1097-0142(19850601)55:11<2709::aid-cncr2820551128>3.0.co;2-a (1985).

56. Hankinson, S. E. & Eliassen, A. H. Endogenous estrogen, testosterone and progesterone levels in relation to breast cancer risk. *The J. Steroid Biochem. Mol. Biol.* **106**, 24–30, DOI: 10.1016/j.jsbmb.2007.05.012 (2007).

57. van den Brandt, P. A. Pooled analysis of prospective cohort studies on height, weight, and breast cancer risk. *Am. J. Epidemiol.* **152**, 514–527, DOI: 10.1093/aje/152.6.514 (2000).

58. Hall, I. J., Newman, B., Millikan, R. C. & Moorman, P. G. Body size and breast cancer risk in black women and white women: The carolina breast cancer study. *Am. J. Epidemiol.* **151**, 754–764, DOI: 10.1093/oxfordjournals.aje.a010275 (2000).

59. Amir, E. Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. *J. Med. Genet.* **40**, 807–814, DOI: 10.1136/jmg.40.11.807 (2003).

60. Results of the ATAC (arimidex, tamoxifen, alone or in combination) trial after completion of 5 years' adjuvant treatment for breast cancer. *The Lancet* **365**, 60–62, DOI: 10.1016/s0140-6736(04)17666-6 (2005).

61. Hallek, M. Chronic lymphocytic leukemia: 2017 update on diagnosis, risk stratification, and treatment. *Am. J. Hematol.* **92**, 946–965, DOI: 10.1002/ajh.24826 (2017).

62. Herishanu, Y. *et al.* Increased serum c-reactive protein levels are associated with shorter survival and development of second cancers in chronic lymphocytic leukemia. *Annals Medicine* **49**, 75–82, DOI: 10.1080/07853890.2016.1232860 (2016).

63. Solans, M. *et al.* Adherence to the western, prudent, and mediterranean dietary patterns and chronic lymphocytic leukemia in the MCC-spain study. *Haematologica* **103**, 1881–1888, DOI: 10.3324/haematol.2018.192526 (2018).

64. Larsson, S. C. & Wolk, A. Overweight and obesity and incidence of leukemia: A meta-analysis of cohort studies. *Int. J. Cancer* **122**, 1418–1421, DOI: 10.1002/ijc.23176 (2007).

65. Brown, L. M. *et al.* Smoking and risk of leukemia. *Am. J. Epidemiol.* **135**, 763–768, DOI: 10.1093/oxfordjournals.aje.a116362 (1992).

66. van Loon, A. J., Goldbohm, R. A. & van den Brandt, P. A. Socioeconomic status and stomach cancer incidence in men: results from the netherlands cohort study. *J. Epidemiol. & Community Heal.* **52**, 166–171, DOI: 10.1136/jech.52.3.166 (1998).

67. Brenner, H., Rothenbacher, D. & Arndt, V. Epidemiology of stomach cancer. In *Methods in Molecular Biology*, 467–477, DOI: 10.1007/978-1-60327-492-0_23 (Humana Press, 2009).

68. Lee, K. R. *et al.* Waist circumference and risk of 23 site-specific cancers: a population-based cohort study of korean adults. *Br. J. Cancer* **119**, 1018–1027, DOI: 10.1038/s41416-018-0214-7 (2018).

69. Zhang, F. F. *et al.* Genetic polymorphisms in alcohol metabolism, alcohol intake and the risk of stomach cancer in warsaw, poland. *Int. J. Cancer* **121**, 2060–2064, DOI: 10.1002/ijc.22973 (2007).

70. Yanada, M. *et al.* Clinical features and outcome of t-lineage acute lymphoblastic leukemia in adults: A low initial white blood cell count, as well as a high count predict decreased survival rates. *Leuk. Res.* **31**, 907–914, DOI: 10.1016/j.leukres.2006.08.004 (2007).

71. Chae, Y. K., Dimou, A., Pierce, S., Kantarjian, H. & Andreeff, M. The effect of calcium channel blockers on the outcome of acute myeloid leukemia. *Leuk. & Lymphoma* **55**, 2822–2829, DOI: 10.3109/10428194.2014.901513 (2014).

72. Molica, S. Sex differences in incidence and outcome of chronic lymphocytic leukemia patients. *Leuk. & Lymphoma* **47**, 1477–1480, DOI: 10.1080/10428190600555819 (2006).

73. Viana, M. B. & Vilela, M. I. O. P. Height deficit during and many years after treatment for acute lymphoblastic leukemia in children: A review. *Pediatr. Blood & Cancer* **50**, 509–516, DOI: 10.1002/pbc.21396 (2007).

74. Coleman, J. F., Theil, K. S., Tubbs, R. R. & Cook, J. R. Diagnostic yield of bone marrow and peripheral blood FISH panel testing in clinically suspected myelodysplastic syndromes and/or acute myeloid leukemia. *Am. J. Clin. Pathol.* **135**, 915–920, DOI: 10.1309/ajcpw10ybrmwswye (2011).