# Cancer survival prediction of six cancer types with germline whole-exome sequencing data from the UK Biobank

Tongqiu Jia

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Public Health

University of Washington

2022

Committee:

Sara Lindström

Ali Shojaie

Program Authorized to Offer Degree:

Epidemiology

University of Washington

## Abstract

Cancer survival prediction of six cancer types with germline whole-exome sequencing data from the UK Biobank

Tongqiu Jia

Chair of the Supervisory Committee:

Sara Lindström

Department of Epidemiology

**Background**: One of the most important tasks in cancer genomics is to predict cancer prognosis based on genetic signatures. We aim to apply machine learning approaches to germline whole-exome sequencing (WES) data to predict survival for breast cancer, colorectal cancer, kidney cancer, lung cancer, lymphoma, and prostate cancer.

**Methods**: We analyzed the UK Biobank exome sequencing data and survival status of 10,721 incident cancer cases. We annotated WES variants with Combined Annotation Dependent Depletion (CADD) and generated gene-level CADD scores, indicating the deleterious mutation impact. Using the gene-level CADD scores, we performed unsupervised feature selection using principal component analysis (PCA), independent component analysis (ICA), random project (RP), autoencoder (AE), denoising autoencoder (DAE), and variational autoencoder (VAE). For each cancer type, we trained logistic regression models to predict cancer survival status using the selected features.

**Results**: All models on all cancer types have a prediction accuracy around 0.5. Overall, the accuracies among deep learning models were comparable to those of linear models. When comparing VAE models with varying latent space, we did not observe an increase in accuracy as the size of latent space increased.

**Conclusion**: Across all six cancer types, the survival prediction accuracy was similar for all models, indicating that more complex deep learning methods did not improve prediction performance. The features or embeddings derived from the six tested dimension reduction methods had limited predictive ability for cancer survival.

# Introduction

The global burden of cancer continues to rise, with an estimated 19.3 million new cancer cases and about 10 million cancer deaths in 2020[1]. As the incidence and mortality increase worldwide, cancer is expected to be the leading cause of death in most countries[2]. While new therapeutic treatments can improve cancer outcomes, the projection of cancer survival remains critical to informing many clinical and personal decisions. One of the most important tasks in cancer prognosis is to provide accurate survival estimation given a patient's clinical and molecular information.

Traditional cancer prognosis prediction relies on clinical presentations, such as cancer type, tumor grade, and stage. Many of these clinical features, including cancer stage, rely on population-level estimates, thus limiting their accuracy when applied at an individual level[3]. The recent advances and cost reduction in next-generation sequencing has propelled cancer research by providing researchers and clinicians with an abundance of multi-omics data, including genomic data, expression data, proteomics data, and epigenetic data. These data enable researchers to assess the contribution of genetic factors to cancer progression and survival, potentially improving individualized prognosis.

There have been large efforts in establishing consortiums to collect, deposit, and share multi-omics and clinical data for research purposes. Notably, the Cancer Genome Atlas (TCGA), Genotype-Tissue Expression (GTEx), the UK Biobank, and All of Us are all recent large initiatives aiming at creating rich resources for the scientific community. Data available through these platforms provide researchers with not only the opportunities to uncover molecular

mechanisms of cancer, but also the resources to develop new methods to better predict cancer survival.

While the unprecedented scale of available multi-omics data is exciting, the high volume and dimensionality pose challenges in identifying true biological signals from noises due to nonlinearity, curse of dimensionality, and associated confounding factors. To deal with these issues, the application of machine learning and deep learning methods on high-dimensional omics data has gained popularity in predicting cancer types and prognosis. A variety of machine learning methods, including Bayesian networks, support vector machines, and decision trees have shown promising results in cancer prognosis prediction[4].

Many machine learning applications have demonstrated success in predicting survival outcomes based on tumor gene expression data. Early attempts include using the expression of 70 genes to separate high and low risk of metastases of breast cancer within five years[5]. Similarly, Beer *et al*. (2002) demonstrated using 50 genes to predict patient survival in early-stage lung adenocarcinomas[6]. With the evergrowing computational power and sample size, deep learning methods in cancer survival prediction has gained popularity in recent years. For example, Kim *et al*. (2020) proposed an unsupervised deep learning model and applied transfer learning to improve cancer survival prediction[7]. Withnell *et al*. (2021) introduced an interpretable variational autoencoder (VAE) based deep learning model for cancer classification tasks[8].

On the other hand, despite that genome-wide association studies (GWAS) have identified numerous disease-related loci, finding germline genetic variants associated with cancer survival has been challenging[9–11]. To date, there has been no application of machine learning methods using exome-wide germline sequencing data to predict cancer survival. In this paper, we aim to

investigate the utility of applying machine learning approaches to germline whole exome-sequencing (WES) data to predict survival for breast cancer, colorectal cancer, kidney cancer, lung cancer, lymphoma, and prostate cancer in the UK Biobank.

# Methods

## Data collection

The study data is collected by the UK Biobank, a large biomedical database containing genetic and health information of half a million participants from the UK. As a large-scale prospective study, the UK Biobank recruited participants living in the UK and aged between 40 and 69 years old. The recruitment period was between 2006 and 2010.

At baseline, participants answered questions related to health and lifestyle, donated blood, urine, and saliva samples, and provided physical measurements. Information about cancer diagnoses and death records are periodically collected by the UK Biobank through linkage to national cancer and death registries.

## Whole-exome sequencing data

UK Biobank has generated whole-exome sequencing (WES) data for the cohort participants, with the first 50,000 individuals released in March 2019, and additional 150,000 individuals released in October 2020. Exomes were captured using a custom version of the IDT xGen Exome Research Panel v.1.0. The sequencing design targets 39 megabases of the human genome, with on average 94.6% of the sites exceeding 20X coverage. The samples were processed by multiplexed paired-end sequencing on the Illumina NovaSeq 6000 platform using

S2 (initial 50k release) and S4 flow cells (all subsequent samples). The raw sequencing data (FASTQs) were aligned to the GRCh38 reference following the OQFE protocol[12]. The sample level variants were called using DeepVariants[13], and the population level variants were jointly aggregated with GLnexus[14].

Selection of study subjects

For this project, study subjects were considered eligible if they 1) consented to participate in the UK Biobank study; 2) had WES data available in the 200k WES data release; 3) had an incident cancer diagnosis, indicated by a valid ICD10 code and/or a record from national cancer registries; and 4) the cancer diagnosis fell into one of the following six cancer types: breast cancer, colorectal cancer, lung cancer, kidney cancer, lymphoma, and prostate cancer. Incident cases were defined as being diagnosed with an invasive cancer after their enrollment date.

Quality control and variant filtering

To ensure the quality of the sample collection and variant call process, we only included samples and variants that passed the following filtering criteria: variants and samples with missingness less than 10%, variants with Hardy-Weinberg equilibrium exact test P-value > $10^{-15}$, variants with minimum read depth (DP) of 7 for both single nucleotide polymorphisms (SNPs) and insertion/deletions (indels), variants with PHRED-scaled quality score (QUAL) no less than 30, and variants with allele balance between 0.15 and 0.85. After applying the filter, 48.5% of variants were removed.

Whole-exome sequencing target region annotation

We annotated the WES target regions with gene names using a two-step consensus approach. The WES target region boundaries were provided by the UK biobank. The first set of

annotations was based on MANE Select[15] exome level annotations (exon-based approach), and the second set of annotations was based on MANE Select gene-level annotations (gene-based approach). Out of the 204,829 capture targets, the exon-based approach annotated 189,273 targets and the gene-based approach annotated 199,874 targets, with a total of 189,185 targets in agreement with each other.

For targets with annotations based on only the gene-based approach, 10,499 targets could be uniquely annotated. There were 43 targets producing multiple annotations to genes within the *UGT1A, PCDHA,* and *CCL3* genes/gene families. To reduce the complexity, we used the gene family name as the final annotation. For 28 targets annotated to multiple genes, we manually verified and curated the gene annotation. The exon-based and gene-based annotations were in disagreement for 53 target regions. The majority of these targets were annotated to unnamed genes. For consistency, we noted all such genes in LOC symbols instead of Ensembl ID. Targets with no annotations for either of the approaches were removed. The final annotations contain 199,859 capture targets that were annotated to 18,662 unique genes.

Defining deleteriousness score for variants and genes

To quantify the impact of each genetic variant, we used Combined Annotation Dependent Depletion (CADD)[16,17], which provides a scoring scheme to indicate the deleteriousness of both SNPs and indels. CADD was trained on 29.4 million observed and simulated variants using an L2 regularized logistic regression model, and then applied to all possible variants. CADD outputs raw scores as well as scaled scores for each of the variants of interest. The raw score has no absolute unit of meaning and is only interpretable on a relative scale – a higher value indicates that a variant is more likely to be deleterious. The scaled score was derived from ranking and “PHRED-scaled” transformation. A variant with a scaled CADD score of 10 is ranked in the top 10% of variants in terms of deleteriousness. Similarly, a variant at the top 1%

has a CADD score of 20, and the top 0.1% has a CADD score of 30. The scaled CADD score has a range from 1 to 99. We annotated our filtered variant set using CADD v1.6 with the scaled scores.

We defined gene boundaries using the annotated WES target regions. To determine gene-level CADD scores for an individual, we assigned the maximum CADD score of all variants within a gene the individual was carrying to be the gene-level CADD score.

Survival outcome and censoring

We considered two types of survival outcomes: all-cause survival and cancer-specific survival. For all-cause survival, the event is recorded death due to any cause. For cancer-specific survival, an event is observed if a person's primary cause of death is due to their primary cancer diagnosis.

The last date of linkage to the death registry was September 30th, 2021. We only considered the primary cause of death and did not use other information, such as secondary causes of death, from the death record.

Model training

We compared six different dimension reduction techniques, including both linear and deep learning methods. All models were trained on the same preprocessed gene-level CADD score as input.

**Principal component analysis (PCA)**: PCA is a linear dimension reduction technique that reduces dimensionality while preserving variance in the original data. PCA obtains a set of orthogonal components that explain the most variance in the data. We used the PCA

implementation from the sklearn package in Python. The output consists of the first 150 principal complements.

**Independent component analysis (ICA)**: ICA is a linear method that transforms the original data into a set of non-Gaussian, independent components. We used the *fastICA* implementation from sklearn package in Python, with 100,000 maximum iterations. The model used 150 components. We fit 10 different ICA models with different random initializations.

**Random projection**: Random projection reduces the data dimensionality by projecting the input on a low dimensional space where components are randomly drawn from a Gaussian distribution. We used the Gaussian random projection function from Python's sklearn package to train 10 times with random initializations.

**Autoencoder (AE)**: AE is an unsupervised neural network that learns a low-dimensional embedding while removing noise from the original data. An AE consists of an encoder that compresses the input data, and a decoder that reconstructs the compressed data to the original space. We implemented the AE model using Tensorflow. Our implementation of the AE model has one hidden layer with 750 latent variables and 0.1 dropout rate. We trained the model using an Adam optimizer with a learning rate of 0.0005 and batch size 100. To account for the model variation between different weight initializations, we repeated the training 10 times with random seed initialization.

**Denoising autoencoder (DAE)**: DAE is a variation of the AE that intentionally adds noise to the input data to prevent overfitting. We randomly generated noise following a normal distribution and added the noise to the input data. Our implementation of the DAE model has one hidden layer with 750 latent variables and 0.1 dropout rate. We repeated training 10 times with random weight initializations, and the training used an Adam optimizer with a learning rate of 0.0005 and batch size 100. The implementation used a Tensorflow backend.

**Variational autoencoder (VAE)**: VAE is a variation of AE with a learned distribution of latent space rather than embedding in the middle (Figure 1). VAE regularizes the distribution of latent space by minimizing the Kullback-Leibler divergence between the learned distribution and a normal distribution. Our VAE models used a mirroring structure of three fully connected layers in both the encoder and decoder. The first two layers of the encoder and decoder used rectified linear unit activation, with a 0.0 dropout rate and match normalization. We trained different models varying the latent space size: 5, 10, 25, 50, 75, and 100. For models with latent space size 5, there were 100 and 25 latent variables in the first two layers, respectively; models with latent space size 10 had 250 and 50 latent variables; the rest of the models had 250 and 100 latent variables. We trained all models using an Adam optimizer with a learning rate of 0.0005 and batch size 50. We implemented the models with Tensorflow.

Performance evaluation

To evaluate the performance of the reduced feature set or learned embeddings, we used the output from each model to predict patients' survival status. We defined survival status as a binary outcome: individuals who died due to any cause, and individuals who were alive at the time of censoring.

As shown in Table 1, when predicting cancer survival as a binary outcome, we have an unbalanced label where each outcome does not correspond to an equal number of cases. To address the label imbalance issue, we subsampled an equal number of cases from each outcome 10 times and used the subsampled dataset as training input.

For binary survival prediction, we used an L1 regularized logistic regression model, with 1000 maximum iterations and *liblinear* solver. We trained model with nested cross-validation: with 20-fold cross-validation splitting the training and testing set, and 5-fold cross-validation to select

regularization coefficients. We implemented the model using the *LogisticRegression* function from Python's sklearn package. All analyses were performed for each cancer type separately.

# Results

Cohort characteristics

We included 10,721 incident cancer cases across six cancer types in our analyses: breast cancer (N=3,256), colorectal cancer (N=1,746), kidney cancer (N=416), lung cancer (N=1,025), lymphoma (N=675), and prostate cancer (N=3,609). The average age of cancer diagnosis was between 64 and 68 years old depending on cancer. Most participants were white based on self-reported ethnic background. Most participants (90%) self-reported no cancer diagnosis at the time of baseline. Breast cancer had the lowest crude proportion of deaths (all-cause death: 7.8%, cancer-specific death: 6.4%), and lung cancer had the highest (all-cause death: 66%, cancer-specific death: 62%). Figure 2 shows the Kaplan-Meier curve of cancer-specific and all-cause survival probability for each cancer type.

Gene-level CADD score does not separate survival outcome

There were total 15,178,802 variants in the UK Biobank provided project-level VCF (pVCF) files. After additional quality control (see Methods), 7,822,596 variants remained. To explore the total impact of deleterious variants exome-wide, we calculated the total gene-level CADD score across all genes for each individual. Figure 3 shows the population distribution of the total exome-wide gene-level CADD score by cancer type. When separating cases by death (yes/no), we did not observe a separation between the two CADD score distributions, comparing those who survived to those who died. These results indicate that the cumulative impact of deleterious variants aggregated at gene level is not predictive of survival status after a cancer diagnosis.

Model performance

The accuracies of cancer survival prediction for each model are shown in Table 2. All models on all cancer types have a prediction accuracy ~0.5, indicating low performance. Overall, the accuracies among deep learning models were comparable to those of linear models, indicating that the more complex models did not improve prediction performance. When comparing VAE models with varying latent space, we did not observe an increase in accuracy as the size of latent space increased.

# Discussion

Across all six cancer types, the survival prediction accuracy was similar for all models, indicating that more complex deep learning methods did not improve prediction performance. The features or embeddings derived from the six tested dimension reduction methods had very limited predictive ability for cancer survival, which does not support our primary hypothesis. We used germline variants generated by WES data as input. While germline genetic variants have been related to cancer heritability previously[18,19], it may not be informative of cancer survival. There are many factors, including non-genetic factors, that can influence cancer prognosis, and our results suggest that a feature set constructed with solely genetic information does not capture the complexity of cancer survival predictors at an individual level. Additionally, while CADD scores quantify the deleteriousness of variants on gene function, they may not be directly relevant to cancer survival.

This work was inspired by DeepProfile[20,21], an ensemble model of variational autoencoder trained from RNA-seq expression data. DeepProfile showed differential accuracies in survival prediction tasks among models, with higher accuracies in more complex models such as VAE

and lower accuracies in linear models such as PCA. Compared to DeepProfile, the key difference in our application is the source of input data. The different distribution between expression data from RNA-seq and the gene-level CADD score of germline variants called from WES could potentially be the main reason for differences in the results. In addition, tissue expression-based models could utilize the numerous open access RNA-seq expression data.

Our study cohort consists of ~10% incident cases who have self-reported cancer diagnosis. Those who have one or more self-reported cancers could be due to incomplete data from cancer registry, reporting errors in self-reported questionnaires, or individuals having prevalent benign cancer diagnoses. Such self-report cancer history among incident cases could potentially introduce noises. Future work could perform additional sensitivity analysis by excluding discordant cases based on self-reported cancer history and cancer registry.

Our results were limited by the sample size of cancer cases, particularly kidney cancer and lymphoma. We initially attempted to predict cancer survival time in a time-to-event model using a Cox proportional hazard model, but the predictive power of such models was minimal (i.e., there was no correlation between the predicted time-to-event and the ground truth). Even when predicting survival status as a binary outcome, the logistic regression model suffered from a reduction in sample size, due to the subsampling technique to address label imbalance. Under the subsampling scheme, the sample size was limited by the class with the lowest number of samples, i.e., observed all-cause deaths. While analysis using cancer-specific deaths could have reduced noises in survival, we were limited by the small sample size, and as a result, we chose to use all-cause deaths. Additionally, we did not have another comparable cancer WES dataset to perform testing in an independent dataset.

Despite the unsatisfactory performance on cancer survival prediction, the embeddings generated from the deep learning models could still be potentially useful for other biologically relevant tasks. For instance, the embeddings could be informative of cancer mutation impact and thus be used to stratify cancer subtypes. Since cancer survival is influenced by complex factors, integrating relevant multi-omics and clinical data could potentially improve survival prediction accuracy. Baek and Lee (2020) proposed a model using RNA expression, WES, and clinical data to predict pancreatic cancer survival and recurrence[22]. Future work could explore these possibilities to further investigate and improve the utility of unsupervised machine learning methods in cancer prognosis and prediction.

# Figures and tables

Table 1. Demographic information of participants

| Characteristic | breast cancer, N = 3,256[1] | colorectal cancer, N = 1,740[1] | kidney cancer, N = 416[1] | lung cancer, N = 1,025[1] | lymphoma, N = 675[1] | prostate cancer, N = 3,609[1] |
|---|---|---|---|---|---|---|
| **Age at recruitment (years)** | 58 (51, 63) | 62 (56, 66) | 61 (56, 65) | 63 (58, 66) | 62 (56, 65) | 63 (59, 66) |
| **Age at primary cancer diagnosis (years)** | 64 (56, 69) | 67 (62, 71) | 67 (62, 71) | 68 (64, 72) | 68 (62, 71) | 68 (64, 72) |
| **Patient sex assigned at birth** | | | | | | |
| Male | 19 (0.6%) | 978 (56%) | 267 (64%) | 541 (53%) | 348 (52%) | 3,609 (100%) |
| Female | 3,237 (99%) | 762 (44%) | 149 (36%) | 484 (47%) | 327 (48%) | 0 (0%) |
| **Self-reported enthnic background** | | | | | | |
| Asian | 63 (1.9%) | 19 (1.1%) | 5 (1.2%) | 10 (1.0%) | 8 (1.2%) | 46 (1.3%) |
| Black | 37 (1.1%) | 22 (1.3%) | 5 (1.2%) | 12 (1.2%) | 4 (0.6%) | 89 (2.5%) |
| Chinese | 10 (0.3%) | 4 (0.2%) | 0 (0%) | 1 (<0.1%) | 1 (0.1%) | 4 (0.1%) |
| Unknown | 17 (0.5%) | 8 (0.5%) | 1 (0.2%) | 9 (0.9%) | 5 (0.7%) | 23 (0.6%) |
| Mixed | 22 (0.7%) | 13 (0.7%) | 2 (0.5%) | 5 (0.5%) | 3 (0.4%) | 18 (0.5%) |
| Other | 26 (0.8%) | 10 (0.6%) | 1 (0.2%) | 8 (0.8%) | 9 (1.3%) | 29 (0.8%) |
| White | 3,081 (95%) | 1,664 (96%) | 402 (97%) | 980 (96%) | 645 (96%) | 3,400 (94%) |
| **Number of self-reported cancers** | | | | | | |
| 0 | 2,855 (88%) | 1,655 (95%) | 391 (94%) | 973 (95%) | 612 (91%) | 3,196 (89%) |
| 1 | 388 (12%) | 78 (4.5%) | 24 (5.8%) | 47 (4.6%) | 61 (9.0%) | 390 (11%) |
| 2 | 13 (0.4%) | 6 (0.3%) | 1 (0.2%) | 5 (0.5%) | 2 (0.3%) | 23 (0.6%) |
| 3 | 0 (0%) | 1 (<0.1%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |
| **All cause death** | 253 (7.8%) | 512 (29%) | 108 (26%) | 676 (66%) | 140 (21%) | 328 (9.1%) |
| **Cancer-specific death** | 207 (6.4%) | 405 (23%) | 95 (23%) | 638 (62%) | 118 (17%) | 234 (6.5%) |

[1] Median (IQR); n (%)

Table 2. Accuracy of cancer survival outcome prediction using features learned from six different models.

| Cancer | Model prediction accuracy (SD[1]) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Linear models | | | Deep learning models | | | | | | | |
| | PCA | ICA | RP | AE | DAE | VAE5L | VAE10L | VAE25L | VAE50L | VAE75L | VAE100L |
| Breast | 0.495 | 0.491 (0.02) | 0.496 (0.02) | 0.484 (0.02) | 0.487 (0.02) | 0.495 (0.02) | 0.500 (0.02) | 0.498 (0.02) | 0.502 (0.03) | 0.501 (0.03) | 0.492 (0.02) |
| Colorectal | 0.495 | 0.495 (0.01) | 0.499 (0.01) | 0.498 (0.02) | 0.494 (0.01) | 0.501 (0.01) | 0.499 (0.02) | 0.499 (0.01) | 0.497 (0.01) | 0.499 (0.01) | 0.494 (0.01) |
| Kidney | 0.487 | 0.485 (0.03) | 0.494 0.03) | 0.493 (0.03) | 0.494 (0.03) | 0.499 (0.03) | 0.492 (0.03) | 0.499 (0.03) | 0.490 (0.03) | 0.503 (0.04) | 0.504 (0.04) |
| Lung | 0.496 | 0.501 (0.02) | 0.495 (0.02) | 0.496 (0.02) | 0.502 (0.02) | 0.499 (0.02) | 0.494 (0.02) | 0.498 (0.02) | 0.500 (0.02) | 0.496 (0.02) | 0.494 (0.02) |
| Lymphoma | 0.516 | 0.514 (0.03) | 0.515 (0.03) | 0.511 (0.04) | 0.516 (0.04) | 0.495 (0.03) | 0.509 (0.03) | 0.509 (0.03) | 0.503 (0.03) | 0.500 (0.03) | 0.499 (0.03) |
| Prostate | 0.502 | 0.511 (0.03) | 0.499 (0.02) | 0.512 (0.02) | 0.512 (0.02) | 0.503 (0.03) | 0.501 (0.02) | 0.502 (0.02) | 0.511 (0.02) | 0.507 (0.02) | 0.503 (0.02) |

[1]SD: standard deviation

Figure 1. Variational autoencoder (VAE) structure.

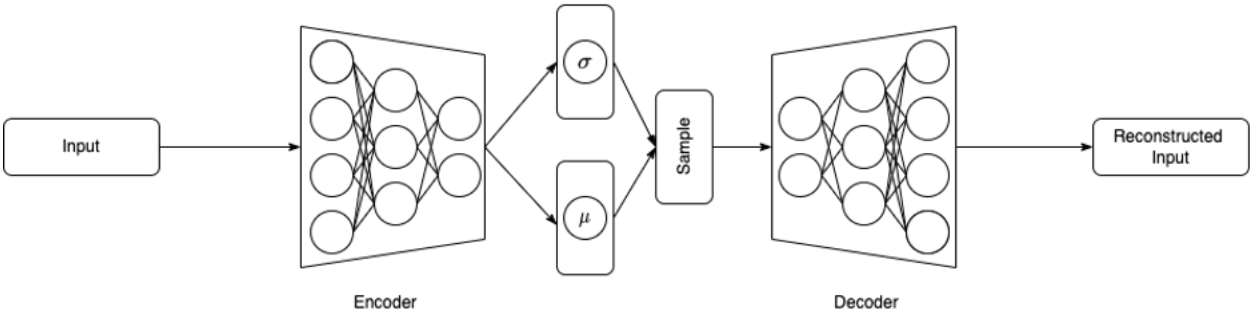


Figure 2. Kaplan-Meier plot of six cancer types based on cancer-specific and all cause survival

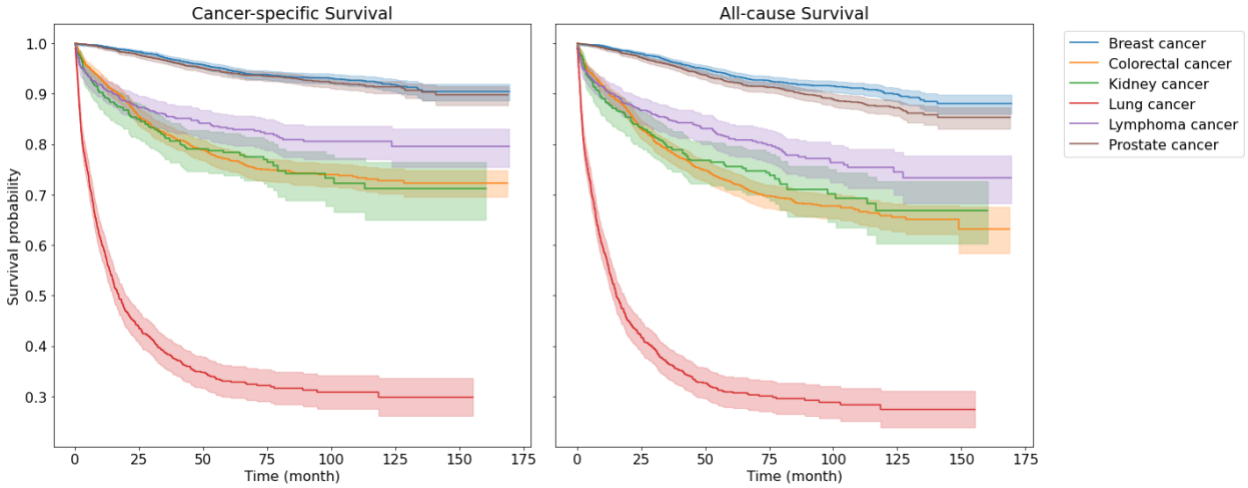
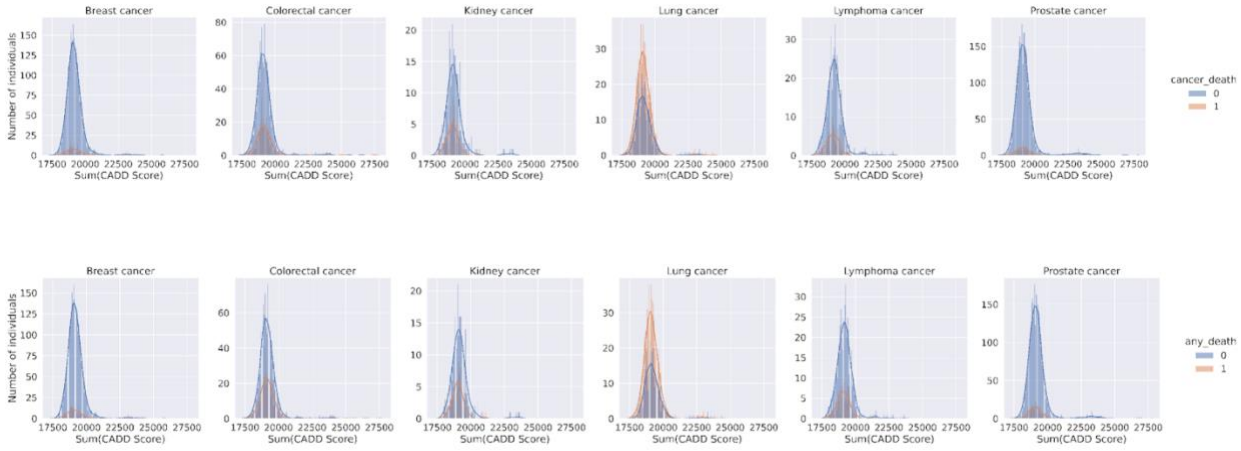


Figure 3. Gene-level CADD score distribution for each cancer type by survival status



# Acknowledgments

This research has been conducted using the UK Biobank Resource, application number 70925. I would like to express my deepest gratitude to my committee members, Dr. Sara Lindström and Dr. Ali Shojaie, who generously provided knowledge and expertise. I am also grateful to my colleagues Austin Hammermeister Suger and Tabitha Harrison, for their feedbacks on the project and moral support.

# References

1. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians* **71**, 209–249 (2021).
2. Bray, F., Laversanne, M., Weiderpass, E. & Soerjomataram, I. The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer* **127**, 3029–3030 (2021).
3. Cheon, S. *et al.* The accuracy of clinicians' predictions of survival in advanced cancer: a review. *Ann Palliat Med* **5**, 22–9 (2016).
4. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. v. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal* **13**, 8–17 (2015).
5. van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
6. Beer, D. G. *et al.* Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine* **8**, 816–824 (2002).
7. Kim, S., Kim, K., Choe, J., Lee, I. & Kang, J. Improved survival analysis by learning shared genomic information from pan-cancer data. *Bioinformatics* **36**, i389–i398 (2020).
8. Withnell, E., Zhang, X., Sun, K. & Guo, Y. XOmiVAE: an interpretable deep learning model for cancer classification using high-dimensional omics data. *Briefings in Bioinformatics* **22**, (2021).
9. Szulkin, R. *et al.* Genome-Wide Association Study of Prostate Cancer–Specific Survival. *Cancer Epidemiology Biomarkers & Prevention* **24**, 1796–1800 (2015).
10. Escala-Garcia, M. *et al.* Genome-wide association study of germline variants and breast cancer-specific mortality. *British Journal of Cancer* **120**, 647–657 (2019).
11. Labadie, J. D. *et al.* Genome-wide association study identifies tumor anatomical site-specific risk variants for colorectal cancer survival. *Scientific Reports* **12**, 127 (2022).
12. Szustakowski, J. D. *et al.* Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nature Genetics* **53**, 942–948 (2021).
13. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology* **36**, 983–987 (2018).
14. Lin, M. F. *et al.* GLnexus: joint variant calling for large cohort sequencing. *bioRxiv* 343970 (2018) doi:10.1101/343970.

15. Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**, 310–315 (2022).
16. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* **46**, 310–315 (2014).
17. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research* **47**, D886–D894 (2019).
18. Carter, H. *et al.* Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer. *Cancer Discovery* **7**, 410–423 (2017).
19. Sayaman, R. W. *et al.* Germline genetic contribution to the immune landscape of cancer. *Immunity* **54**, 367-386.e8 (2021).
20. Dincer, A. B., Celik, S., Hiranuma, N. & Lee, S.-I. DeepProfile: Deep learning of cancer molecular profiles for precision medicine. *bioRxiv* 278739 (2018) doi:10.1101/278739.
21. Dincer, A. B. *et al.* A deep profile of gene expression across 18 human cancers. *Manuscript submitted for publication*.
22. Baek, B. & Lee, H. Prediction of survival and recurrence in patients with pancreatic cancer by integrating multi-omics data. *Scientific Reports* **10**, 18951 (2020).