



## Article

# A Study on Survival Analysis Methods Using Neural Network to Prevent Cancers

Chul-Young Bae <sup>1</sup>, Bo-Seon Kim <sup>1</sup>, Sun-Ha Jee <sup>2</sup>, Jong-Hoon Lee <sup>3</sup>  and Ngoc-Dung Nguyen <sup>3,\*</sup> <sup>1</sup> Mediage Research Center, Seongnam-si 13449, Republic of Korea<sup>2</sup> Department of Epidemiology and Health Promotion, Institute for Health Promotion, Graduate School of Public Health, Yonsei University, Seoul 03722, Republic of Korea<sup>3</sup> Moadata AI Labs, Seongnam-si 13449, Republic of Korea\* Correspondence: [nguyen@moadata.ai](mailto:nguyen@moadata.ai); Tel.: +82-70-4099-5131

**Simple Summary:** According to cancer statistics published in 2020, there were 19.3 million new cancer cases and almost 10.0 million cancer deaths worldwide. This suggests that cancer is one of the main health threats worldwide. Survival analysis, combining the exponential distribution with regression analysis, can predict the time when a specific event will occur. This nationwide follow-up study aims to present survival deep learning models to assist in the early identification of cancer incidence. Our models consistently achieved high performance in 10 types of cancer. By applying the techniques outlined in this paper, clinical biomarkers, demographic, and anthropometric data can be utilized to predict the risk of cancer occurrence.

**Abstract: Background:** Cancer is one of the main global health threats. Early personalized prediction of cancer incidence is crucial for the population at risk. This study introduces a novel cancer prediction model based on modern recurrent survival deep learning algorithms. **Methods:** The study includes 160,407 participants from the blood-based cohort of the Korea Cancer Prevention Research-II Biobank, which has been ongoing since 2004. Data linkages were designed to ensure anonymity, and data collection was carried out through nationwide medical examinations. Predictive performance on ten cancer sites, evaluated using the concordance index (c-index), was compared among nDeep and its multitask variation, Cox proportional hazard (PH) regression, DeepSurv, and DeepHit. **Results:** Our models consistently achieved a c-index of over 0.8 for all ten cancers, with a peak of 0.8922 for lung cancer. They outperformed Cox PH regression and other survival deep neural networks. **Conclusion:** This study presents a survival deep learning model that demonstrates the highest predictive performance on censored health dataset, to the best of our knowledge. In the future, we plan to investigate the causal relationship between explanatory variables and cancer to reduce cancer incidence and mortality.

**Keywords:** survival analysis; recurrent neural network; LSTM; Cox PH; biomarkers; cancer; cohort; follow-up



**Citation:** Bae, C.-Y.; Kim, B.-S.; Jee, S.-H.; Lee, J.-H.; Nguyen, N.-D. A Study on Survival Analysis Methods Using Neural Network to Prevent Cancers. *Cancers* **2023**, *15*, 4757. <https://doi.org/10.3390/cancers15194757>

Academic Editors: Teng Huang, Qiong Wang and Yan Pang

Received: 22 August 2023

Revised: 15 September 2023

Accepted: 20 September 2023

Published: 27 September 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Cancer stands as a leading cause of premature death in upper-middle- and high-income countries [1]. Specifically, female breast, lung, colorectal, prostate, stomach, and liver cancers account for the highest incidence and mortality rates [1,2]. Early personalized prediction of cancer incidence holds paramount importance for the population at risk. Substantial evidence regarding the risk factors of noncommunicable diseases, including cancer, has primarily been drawn from large-scale follow-up studies, with many of them employing survival analysis.

Survival analysis is centered on determining the probability of surpassing a certain period without encountering the event, as well as relevant outcomes. It has demonstrated

its robustness in comparing risks among subgroups within the targeted population and has found applications in diverse fields, spanning from economics to healthcare [3]. Consequently, survival analysis has been utilized in the quest to identify factors associated with cancer and predict cancer survival [4].

The Cox proportional hazard (Cox PH) regression model [5] has emerged as a popular tool within the realm of survival analysis due to its interpretability and predictive ability. However, it rests upon strong assumptions regarding covariates. To mitigate these limitations, numerous models have been developed based on the Cox PH model, incorporating artificial intelligence techniques [4]. One such machine learning approach, the random survival forest, has gained prominence. Certain studies [6–8] were among the first to introduce a single hidden layer network for survival analysis. Building upon similar concepts, Katzman [9] further developed a deep survival network for personalized treatment recommendations. The predictive efficacy of survival networks was markedly enhanced by the application of multitask learning and residual connections, as proposed in the model known as DeepHit [10].

Recent research has suggested that survival neural networks outperform their machine learning and traditional statistical counterparts [10,11]. Leveraging data from a large-scale blood-based cohort, we conducted a comparative assessment of the performance of various survival deep networks, using Cox PH as the benchmark, to predict ten cancers that rank among the most common globally [2,12]. Furthermore, we introduced nDeep—a model that, to the best of our knowledge, achieves the highest possible performance on real censored health datasets. In an attempt to interpret the effects of those biomarkers in each model, feature importance was also calculated.

## 2. Materials and Methods

### 2.1. Subjects

A total of 160,407 observations (95,229 men and 62,169 women) were collected from a blood-based cohort within The Korean Cancer Prevention Study-II Biobank, an ongoing study initiated in 2004 through nationwide medical examinations [13]. Participants underwent annual follow-ups utilizing personal identification numbers, enabling linkage with the National Cancer Center registry, hospital admission records, and death registers. The integrity of all linkages was guaranteed with anonymity safeguards. During routine health check-ups, individuals diagnosed with cancer or other severe conditions (such as organ failure, diabetes, malignant hypertension), as well as those who died, were subsequently excluded from the study.

### 2.2. Characteristics, Biomarkers, and Events

Anthropometry (height and weight) was recorded by InBody (Biospace, Cheonan-si, Korea) when participants wore light clothes. For waist circumference, the thinnest area between the iliac crest and the inferior part of the lowest rib was measured in an upright position. Forced vital capacity and forced expiratory volume in 1 second were recorded by an electronic spirometer in standing position, while the better results out of two records were taken.

Seated blood pressure was then taken by a registered nurse or technician using a sphygmomanometer after resting 5 min. Fasting blood glucose, total cholesterol, and other biomarkers were measured in the laboratory of the hospital by a COBAS INTEGRA 800 (Roche, Berlin, Germany) and a 7600 Analyzer (Hitachi, Tokyo, Japan). More details can be found at [13]. For our final analysis, these features were included as covariates:

- Demographic: AGE—age, years; SEX—sex, 1 as male and 2 as female.
- Metabolic characteristics: WT—weight, kilograms; HT—height, centimeters; and BMI—body mass index =  $\text{weight}/\text{height}^2$  ( $\text{kg}/\text{m}^2$ ); WC—waist circumference, centimeters; SBP—systolic blood pressure, mmHg; DBP—diastolic blood pressure, mmHg; CHO total cholesterol, mg/dL; HDL—high-density cholesterol, mg/dL; TG—triglycerides,

mg/dL; LDL—low-density cholesterol, mg/dL; GLU—glucose, mg/dL; PP—pulse pressure = systolic blood pressure – diastolic blood pressure, mmHg.

- Liver function: ALB—albumin, g/dL; GLOBULIN—globulin, g/dL; AGR—albumin globulin rate; BIL—total bilirubin, mg/dL; DBIL—direct bilirubin, mg/dL; ALP—alkaline phosphatase, units/L; AST—aspartate aminotransferase, units/L; ALT—alanine aminotransferase, units/L; GGTP—gamma-glutamyl transpeptidase, units/L.
- Kidney function: CREAT—creatinine, mg/dL; BUN—blood urea nitrogen, mg/dL; SG—specific gravity; PH—urine pH; CCR—creatinine clearance rate =  $(140 - \text{Age}) \times \text{Weight} / (\text{Creatinine} \times 72)$  for Men OR  $((140 - \text{Age}) \times \text{Weight} / (\text{Creatinine} \times 72)) \times 0.85$  for Women.
- Pancreas function: AMYLASE—amylase, units/L.
- Pulmonary function: FVC—forced lung capacity, measured in liters; FEV1—forced expiratory volume, measured in liters.

The study sample is representational regarding age. The number of men accounts for 60% of the total sample (Table 1).

**Table 1.** Basic information on demography, anthropometry, and clinical characteristics of study population.

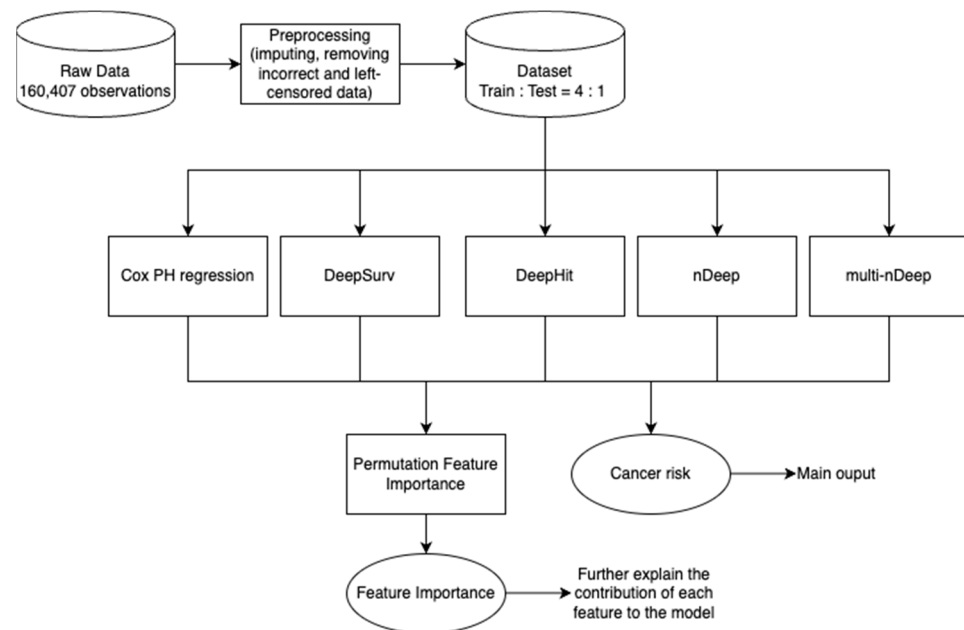
	Mean ± SD	Min	Max	Male Mean ± SD	Female Mean ± SD
AGE	41.62 ± 10.53	9.00	88.00	42.04 ± 9.95	40.98 ± 11.31
HT	166.80 ± 8.40	130.00	198.00	171.75 ± 5.95	159.31 ± 5.52
WT	65.79 ± 12.00	32.00	137.00	72.12 ± 9.93	56.23 ± 7.78
WC	80.58 ± 9.45	41.00	127.00	84.80 ± 7.65	74.21 ± 8.26
BMI	23.53 ± 3.17	12.91	42.36	24.42 ± 2.88	22.18 ± 3.10
SBP	117.65 ± 14.11	61.00	199.00	121.23 ± 12.87	112.25 ± 14.19
DBP	74.12 ± 10.04	30.00	134.00	76.55 ± 9.56	70.45 ± 9.63
CHO	188.83 ± 33.02	3.68	382.00	192.04 ± 32.86	183.99 ± 32.68
HDL	52.27 ± 10.70	2.00	118.00	48.80 ± 9.01	57.50 ± 10.92
TG	132.82 ± 83.35	30.00	690.00	155.29 ± 90.44	98.96 ± 56.51
LDL	112.13 ± 31.03	1.00	299.60	115.23 ± 31.25	107.47 ± 30.11
FVC	26.93 ± 30.82	0.04	158.00	25.33 ± 30.39	29.16 ± 31.33
FEV1	28.64 ± 33.96	0.42	178.00	26.58 ± 33.25	31.56 ± 34.81
ALB	4.53 ± 0.25	2.20	5.90	4.58 ± 0.24	4.45 ± 0.24
GLOBULIN	2.76 ± 0.17	1.80	3.80	2.74 ± 0.17	2.78 ± 0.16
AGR	1.64 ± 0.13	0.85	2.39	1.67 ± 0.13	1.59 ± 0.11
BIL	0.87 ± 0.33	0.10	2.93	0.94 ± 0.34	0.75 ± 0.28
DBIL	0.33 ± 0.12	0.04	1.07	0.36 ± 0.12	0.28 ± 0.10
ALP	133.52 ± 50.56	10.00	445.00	142.70 ± 50.77	119.69 ± 46.93
AST	22.69 ± 8.90	1.00	135.00	24.65 ± 9.55	19.75 ± 6.85
ALT	24.43 ± 17.23	1.00	180.00	29.46 ± 18.95	16.87 ± 10.68
GGTP	34.23 ± 32.45	1.90	390.00	44.58 ± 36.75	18.67 ± 14.83
GLU	90.56 ± 15.19	10.00	208.00	92.37 ± 16.18	87.83 ± 13.07
AMYLASE	70.98 ± 19.00	3.00	195.00	69.63 ± 18.65	73.04 ± 19.35
CREAT	0.98 ± 0.18	0.38	2.40	1.09 ± 0.14	0.83 ± 0.12
BUN	13.74 ± 3.25	0.45	59.00	14.50 ± 3.13	12.57 ± 3.07
SG	1.02 ± 0.01	1.00	1.05	1.02 ± 0.01	1.02 ± 0.01
PH	5.68 ± 0.76	0.50	9.00	5.66 ± 0.75	5.72 ± 0.77
PP	43.54 ± 9.64	0.00	101.00	44.69 ± 9.50	41.80 ± 9.59
CCR	92.92 ± 20.02	20.14	213.37	91.66 ± 19.86	94.85 ± 20.12

Events of interest are ten cancer sites that are among the cancers with top incidence and top mortality rates: thyroid, gastric, breast, colorectal, lung, prostate, liver, kidney, uterine-cervical, and lymphoma cancers.

### 2.3. Project Pipeline

For the raw dataset, observations that are left-censored or contain values that are 6SD away from the mean were excluded (3.3% of the sample). Features with a missing rate of less than 25% were retained and imputed. The time variable was defined by subtracting the year of study enrolment from the year of getting diagnosed with any of the above cancers (event case) or finishing/quitting the study before any event (censored case). The processed data were then fed into each of the five models so that cancer risk

and feature importance could be calculated (Figure 1). Our analysis process are released at <https://github.com/ngocdung03/nDeep>.



**Figure 1.** Project pipeline.

#### 2.4. Model Description

Four neural networks and one Cox PH model were included for predictive performance comparison.

##### 2.4.1. Cox PH Regression

Developed in 1972, Cox PH is a semi-parametric regression model for determining the association between predictors (covariates) and the rate of an event happening at a specific time point (hazard rate).

$$h(t) = h_0(t) \exp(\beta \times x_i)$$

where,

- $t$  is the survival time
- $h(t)$  is the hazard function
- $x_i$  is the vector of covariates
- $h_0(t)$  is the baseline hazard when all the  $x_i$  are equal to zero.
- $\beta$  is the vector of coefficients of  $x_i$

By maximizing this partial likelihood, the vector of  $\beta$  is estimated:

$$L_c(\beta) = \prod_{i: E_i=1} \frac{\exp(\beta \times x_i)}{\sum_{j \in R(T_i)} \exp(\beta \times x_j)}$$

where,

- $E_i$ ,  $T_i$ , and  $x_i$  are event indicator, survival time, and baseline covariates of the  $i$ th observation.
- The likelihood is defined in non-censored observations ( $E_i = 1$ ).
- $R(t) = \{i: T_i \geq t\}$  is the set of participants at risk of event at time  $t$ .

##### 2.4.2. DeepSurv

DeepSurv is a deep neural network (DNN) that outputs the estimate of hazard function  $\hat{h}_\theta(x)$  with  $\theta$  as the weights of the network [9]. Unlike the model of [8], it applied modern

techniques for training DNN, including weight decay regularization, batch normalization, Rectified Linear Units (ReLU), dropout, and gradient descent optimization.

#### 2.4.3. DeepHit

DeepHit is a multitask neural network that is formed by shared layer(s) and K task-specific sub-networks (K is the number of tasks). There is a residual connection from input to each task-specific sub-network. Deep residual learning [14] is a technique for overcoming the accuracy degradation when increasing the depth of the network. In addition, with multitask learning, DeepHit can alleviate the competing risk issue, which is common in survival analysis and studies of cancer.

In this study, we performed two multitask learning DeepHit models on these groups of tasks: Group 1—thyroid, gastric, breast, colorectal, and lung cancers; and Group 2—prostate, liver, kidney, uterus-cervical, and lymphoma cancers.

#### 2.4.4. nDeep

This model contains three layers of long short-term memory (LSTM) and one fully connected layer that outputs predicted hazard. LSTM is a type of recurrent network (RNN) that has feedback connections that can address vanishing or exploding gradient issues of RNN [15]. Its architecture transforms survival data into sequences of n time steps, which can then be processed and analyzed. A basic LSTM unit includes a cell, an input gate, an output gate, and a forget gate (Figure 2). The equations for a forward pass of an LSTM cell are as follows:

$$\begin{aligned} f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \\ i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \\ o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \\ \tilde{c}_t &= \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\ h_t &= o_t \odot \sigma_h(c_t) \end{aligned}$$

where,

- The initial values  $c_0 = 0$  and  $h_0 = 0$
- $x_t \in R^d$ : vector of input
- $f_t \in (0, 1)^h$ : vector of forget gate's activation
- $i_t \in (0, 1)^h$ : vector of input/update gate's activation
- $o_t \in (0, 1)^h$ : vector of output gate's activation
- $h_t \in (-1, 1)^h$ : vector of hidden state, also output
- $\tilde{c}_t \in (-1, 1)^h$ : vector of cell input activation
- $c_t \in R^h$ : vector of cell state
- $W \in R^{h \times d}$ ,  $U \in R^{h \times d}$  and  $b \in R^h$ : vectors of weight matrices and bias
- $\odot$ : element-wise product (Hadamard product)

#### 2.4.5. Multi-nDeep

Being inspired by the combination of multitask learning and residual connection, multi-nDeep has a shared block of fully connected layers and K task-specific blocks of nDeep network (Figure 3). The residual is connected from the input to each task-specific block. Models of multitask-nDeep were trained on these groups of tasks: Group 1—thyroid and gastric; Group 2—breast, colorectal, and lung; Group 3—prostate and liver; and Group 4—kidney, uterus-cervical, and lymphoma cancers.

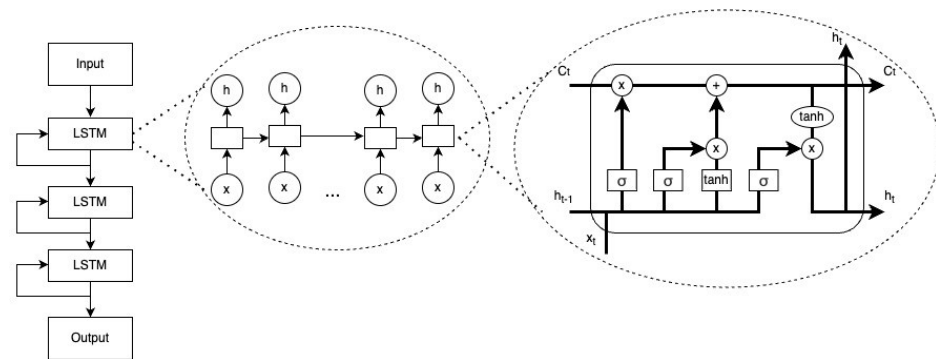


Figure 2. Architecture of nDeep.

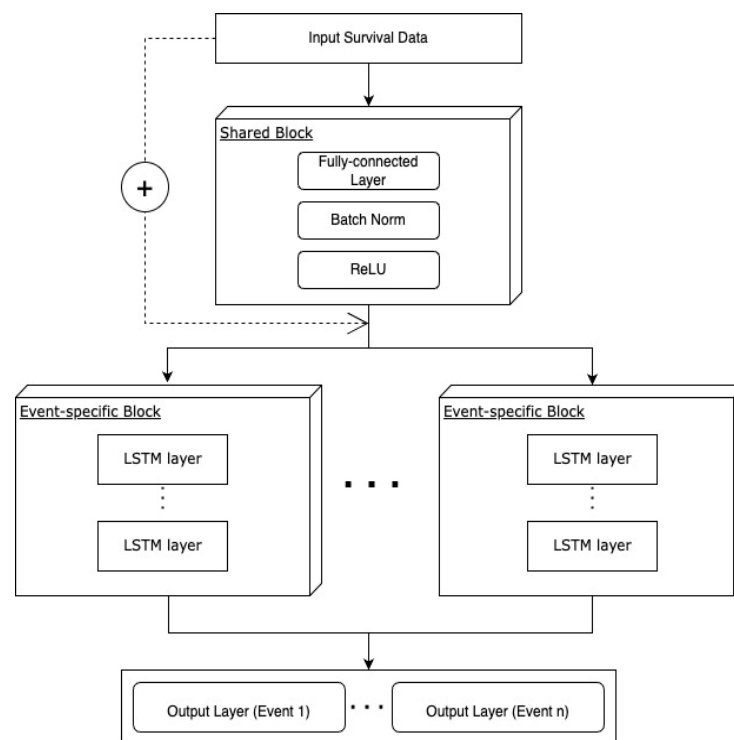


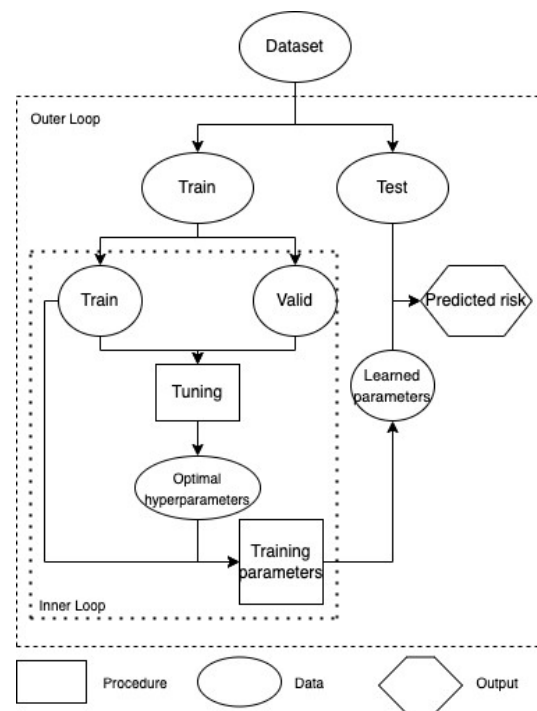
Figure 3. Architecture of mtl-nDeep.

### 2.4.6. Hyperparameters of DNN Models

Hyperparameters of each DNN model are shown in Supplementary Table S1. Their optimization progress is depicted in Supplementary Figure S1.

### 2.5. Model Learning

Four networks were involved in two training loops. The outer loop started with the dataset being randomly divided into a train set and a test set (ratio 4:1). In the inner loop, the valid set was further sampled on the train set with the probability of 0.2 so that hyperparameters were optimized by an automatic framework called Optuna with 60 iterations. This framework employs a Tree-structured Parzen Estimator (TPE), a state-of-the-art algorithm based on Bayesian optimization, which is faster and more efficient than grid search (Supplementary Figure S1). The tuned hyperparameters were used for training parameters of each neural network. The inner loop returned a learned model, which was run on the test set 30 times to output predicted cancer hazard risks and averaged c-index (Figure 4).



**Figure 4.** The workflow for optimizing hyperparameters of the models. The inner loop is a process to find the optimal hyperparameters during a training process. Furthermore, the outer loop is a process to evaluate the optimal hyperparameters that were obtained by training with test sets once again.

The loss function used for each DNN is calculated by the sum of L1 + L2 where: L1 is the negative log partial likelihood:

$$L1 = -\sum_{i=1}^N I_i \left[ \hat{h}_\theta(x_i) - \log \sum_{j \in R(t_i)} e^{\hat{h}_\theta(x_j)} \right]$$

where,

- $N$ : number of subjects
- $I$ : indicator function (1: death, 0: otherwise)
- $\theta$ : weights of the network
- $\hat{h}_\theta$ : estimated hazard function
- $x_i, x_j$ : vectors of covariates
- $R(t_i)$ : set of subjects at risk at time  $t_i$

L2 is the ranking loss function:

$$L2 = \sum_{i \neq j} I_{i,j} \cdot \eta \left( \hat{F}(s^{(i)} | x^{(i)}), \hat{F}(s^{(j)} | x^{(j)}) \right)$$

where,

- $I_{i,j}$ : indicator function of acceptable pairs  $(i,j)$  [16]
- $\eta(x, y) = \exp\left(\frac{-(x-y)}{\sigma}\right)$ : convex loss function
- $\hat{F}$ : estimated cumulative incidence function
- $s$ : time
- $x$ : covariate

### 2.6. Performance Metric

Model accuracy was evaluated by Harrell’s concordance index (c-index) [16]. A c-index of 1 represents perfect accuracy. Contrary, a c-index which is 0.5 or less suggests



the model performs no better than a naïve model. Five-fold cross-validation was performed 30 times for each task, with the test set being sampled with replacement. The final c-index was achieved by averaging the c-indices of those 30 trials.

### 2.7. Feature Importance

For each model, permutation feature importance (PFI) was calculated by the module permutation importance of *sklearn* version 1.0.2—a Python library. PFI scores were evaluated based on the c-index. The values of PFI are presented in figures and grouped by a separation line for interpretation. The groups of features represent demographic, metabolic, liver function, pulmonary function, kidney function, and pancreas function.

### 2.8. Ethical Considerations

Informed consent of participants was received under the support of the Metabolic Syndrome Research Program of the Seoul Metropolitan Government in December 2005. This study was approved by the Yonsei University Institutional Review Board (IRB Approval No. 4-2011-0277).

The current study is approved by the Korea Institute of Bioethics Policy Electromagnetic Concern Committee to be exempt from examination. All methods were used based on ethical regulations and guidelines.

## 3. Results

Table 2 indicates the number of cases on each cancer site and the corresponding rate over the total number of samples. Cases/total sample rates were small, suggesting an imbalance in data.

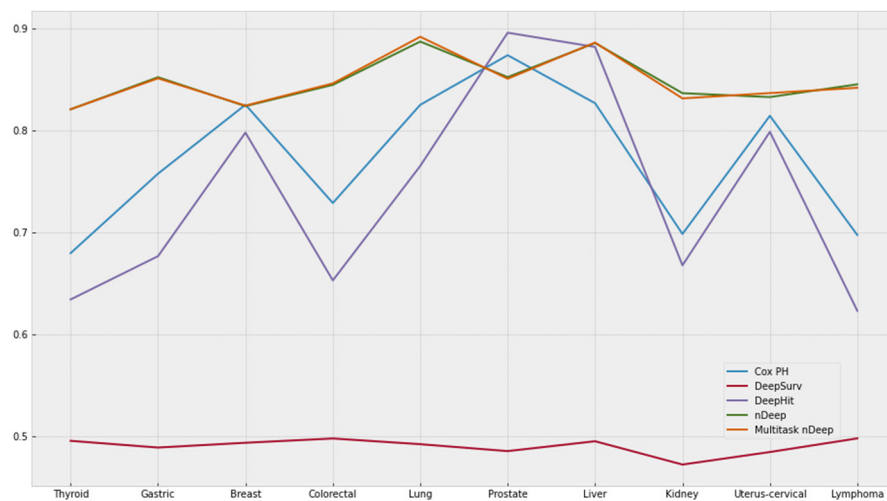
**Table 2.** Number and rate of cancer cases on each site.

Cancer Site	Number of Cases	Cases/Total Cancer Cases	Cases/Total Sample
Thyroid	3385	0.33120	0.02189
Gastric	1499	0.15109	0.00969
Breast	1169	0.11783	0.00756
Colorectum	1031	0.10392	0.00667
Lung	757	0.07630	0.00489
Prostate	697	0.07026	0.00451
Liver	439	0.04425	0.00284
Kidney	352	0.03548	0.00228
Uterus-cervical	306	0.03084	0.00198
Lymphoma	286	0.02883	0.00185

### 3.1. Comparing c-Index of Each Model

Figure 5 shows the c-index of five models on different tasks of predicting cancer. DeepHit and Cox PH regression models attained the equivalent pattern of c-index. Although c-indices of DeepHit on many cancers are slightly lower as compared to Cox PH, DeepHit, a multitasking model, reached c-index as high as 0.8962 and 0.8822 for prostate and liver cancers. DeepSurv did not indicate any predictive power with the c-index of about 0.5 for every model. On the other hand, nDeep and multi-nDeep achieved the c-indices, which are consistently more than 0.8. Although there was a slight improvement in multi-nDeep as compared to nDeep, the effects were not remarkable. In Supplementary Table S2, the best c-index of each cancer is presented in bold. For reference purpose, c-indices of variants of Cox PH were included in Table S3.





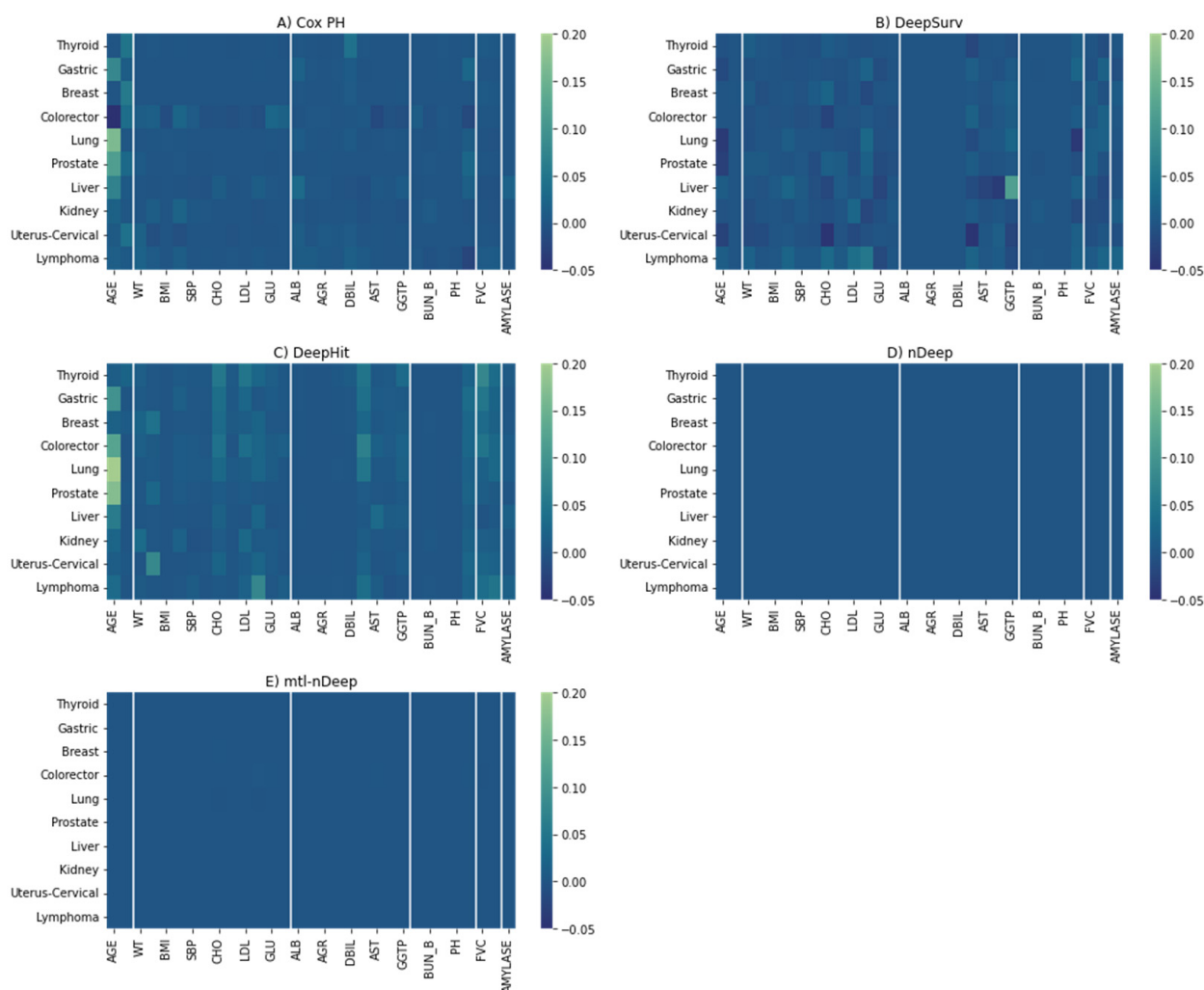
**Figure 5.** Best c-index of each model. The y-axis represents c-index values.

### 3.2. Feature Importance

For Cox PH, the most important features are demographic ones (Age and Sex), followed by liver function (Figure 6A). Age is the most decisive factor in the majority of Cox PH models, while sex applies to thyroid cancer, breast cancer, colorectal, and uterus-cervical cancers (particularly, women have a higher risk of these cancers as compared with men in this study). In addition, direct bilirubin, a liver function metric, was also a crucial determinant of thyroid cancer. Metabolic characteristics even showed higher importance than demographic factors for kidney and lymphoma cancers, where waist circumference and weight were the main predictors. Kidney function characteristics showed moderate importance for some cancers (creatinine clearance rate for gastric, prostate, and liver cancers, and blood urea nitrogen for kidney cancer). Amylase, a pancreas function index, and creatinine clearance rate were moderately important for liver cancer.

Similar patterns were observed for DeepHit on gastric, lung, prostate, liver, kidney, and lymphoma cancers, although there were more contributions from factors across categories other than the dominant factor (Figure 6C). For kidney and lymphoma cancer, metabolic characteristics were still the most important but attributed to other metrics (weight and low-density lipoprotein, and triglyceride, respectively). Regarding colorectal cancer, age is the main determinant rather than sex. PFIs, according to thyroid and breast cancers, are relatively equitably distributed across categories. Uterus-cervical cancer is mostly attributed to height.

The figure of feature importance for DeepSurv (Figure 6B) was almost uninformative, with rather random distributions around zero. In Figure 6D,E, with PFI score  $<0.001$ , the effect of each feature on nDeep models was minor. Generally, the patterns of multitask-nDeep were similar to nDeep.



**Figure 6.** The heatmaps of PFI for five models. Differences in the c-index after removing a feature were represented in a manner where the brighter the color, the more important the feature is. (A) Cox PH. (B) DeepSurv. (C) DeepHit. (D) nDeep. (E) mtl-nDeep.

#### 4. Discussion

Only two models, nDeep and multitask nDeep, outperformed the traditional Cox PH regression model. On the other hand, DeepSurv exhibited no improvement over a naïve model. This is inconsistent with the results presented in reference [9], where the c-index of DeepSurv can be more than 0.7. Similarly, DeepHit showed comparable c-indices with Cox PH in our study, while its performance was significantly better in generated data [10]. That signifies the disparities in nature between the simulated and real censored data. One of the persistent challenges in follow-up health data is the low rate of uncensored cases, especially when the event is a rare disease such as cancer, which poses significant hurdles for the accuracy of survival analysis [17]. Furthermore, in our study, DeepSurv is the only neural network without residual connections, which could explain its underperformance. In contrast, other than the well-known Cox PH, nDeep emerges as the best model in terms of both accuracy and cost-effectiveness in the current study. It seems that the absence of a particular feature did not hinder the performance of nDeep and its variants.

There are two main factors that affect nDeep performance. Firstly, drawing inspiration from natural language processing, we treated the observed data for a participant and a visit as a single sequence with  $n$  time steps and one subsequence with  $m$  time steps (where  $m < n$ ), respectively. Secondly, our model architectures incorporated the benefits of residual connections and multiple gates in an LSTM memory cell to effectively capture

the dynamic nature of covariates. Moreover, finding the optimal models was eased by applying techniques including TPE algorithm.

The connection between age and cancer is well-established [18]. In fact, our analysis based on Cox PH regression models indicates that age has an impact on nearly all types of cancers. Aging decreases the effectiveness of autophagy, a process responsible for damaged organelles removal, leading to uncontrollable cellular homeostasis and cancer genesis [19]. Some cancers, such as breast and uterus-cervical cancers, are strongly affected by sex, which is expected as these cancers predominantly or exclusively affect females [2,12]. Additionally, our study reveals a higher prevalence of colorectal cancer in women compared to men. This is possibly due to both genetic and environmental factors, including dietary habits [20]. Similarly, the female predominance in thyroid cancer cases can be attributed to factors such as sex hormones, genetics, and the immune system [21–24].

Neural network like DeepHit helps learn the non-linear effects of factors beyond age and sex in cancer risk assessment. It was observed that metabolic metrics, liver function, pulmonary function, and kidney function indices serve as predictors for certain cancer types, including thyroid, gastric, breast, colorectal, kidney, and lymphoma cancers.

Obesity-related metabolic disorders have long been recognized as risk factors for thyroid cancer [25]. Specifically, WC is a strong indicator of the association between overweight/obesity and thyroid cancer [26]. Obesity also highly promotes the progression of various other cancers through mechanisms involving hormones, insulin and insulin-like growth factors, sex steroids, and obese inflammation [27,28]. Increased inflammation in adipose tissue, a consequence of obesity, results in changes in adipokine secretion patterns and the release of pro-inflammatory cytokines, leading to insulin resistance rises in tissues with active metabolism [29]. Overexpression of insulin-like growth factors may have neoplastic effects by promoting cell cycle progression and inhibiting apoptosis, either directly or indirectly, through interaction with established oncogenic systems such as steroid hormones and integrins [30]. Moreover, excess peripheral adipose tissue contributes to steroid aromatization, leading to elevated estrogen levels. Androgens and androgenic antecedents are changed over to estradiol by the chemical aromatase. Within the setting of weight and abundance of fat tissue, aromatase expanded action leads to higher transformation rates coming about in higher levels of estrogens [31]. Estrogen advances tumorigenesis in endometrial tissue by incitement of cell multiplication and hindrance of apoptosis [32].

A recent study demonstrated a link between metabolic dysfunction-associated fatty liver and thyroid cancer, even after adjusting for WC [33]. The relationship between thyroid cancer, a hormone-sensitive cancer, and metabolic dysfunction was found to be modified by menopausal status in females. Higher levels of liver function markers were pointed out to be associated with a reduced risk of colorectal cancer [34]. In the ileum, bilirubin undergoes deconjugation through bacterial and mucosal processes, potentially possessing antioxidant and anti-inflammatory properties [35]. Considering that inflammation and oxidative stress are pivotal contributors to cancer development, this association is noteworthy [36]. The link between pulmonary function and cancer is primarily limited to lung cancer [37]. Likewise, kidney function has been found to be associated with kidney cancer mainly.

This study represents the first attempt to compare different neural network models and traditional models in survival analysis using clinical biomarkers on a large-scale cohort. Our findings align with prior evidence indicating that deep neural networks outperform other methods in healthcare survival analysis. Furthermore, we have clarified the role of different factors in predicting cancer incidence. However, according to the PFI heatmaps, while Cox PH regression heavily relies on information regarding sex and age, cancers appear to be more evenly affected by all the features in our models. Therefore, in the future, we should investigate advanced models capable of distinguishing key factors clearly and incorporate additional biomarkers for more accurate cancer prediction. Additionally, exploring the causal relationship between key factors and specific cancers is crucial, as it can help identify causal factors and facilitate cancer prevention efforts.

## 5. Conclusions

We have introduced nDeep and its multitasking version for predicting cancer risks based on clinical biomarkers. These algorithms transform health follow-up data into sequences that can be fed into recurrent neural networks, enabling high predictive performance. Our model can assist health care providers in determining people at risk of ten different cancers.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/cancers15194757/s1>, Figure S1: Hyperparameters optimization of 4 models based on TPE algorithms; Table S1: Hyperparameters of DNN models; Table S2: C-index of each model and imputation method; Table S3: C-index of variants of Cox PH models.

**Author Contributions:** C.-Y.B. conceptualized the study, designed the research. B.-S.K. performed the literature review and critically revised the article for important intellectual content. S.-H.J. supervised the study. J.-H.L. conducted the data collection and analysis. N.-D.N. interpreted the results and drafted the manuscript. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** The current study is approved by the Korea Institute of Bioethics Policy Electro-magnetic Concern Committee to be exempt from examination. All methods were conducted based on ethical regulations and guidelines.

**Informed Consent Statement:** Informed consents of participants were received under the support of the Metabolic Syndrome Research Program of the Seoul Metropolitan Government in December 2005. This study was approved by the Yonsei University Institutional Review Board (IRB Approval No. 4-2011-0277).

**Data Availability Statement:** The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- World Health Organization—The Top 10 Causes of Death. Available online: <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death> (accessed on 2 December 2022).
- Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: {GLOBOCAN} estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA A Cancer J. Clin.* **2021**, *71*, 209–249. [[CrossRef](#)]
- Emmert-Streib, F.; Dehmer, M. Introduction to survival analysis in practice. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 1013–1038. [[CrossRef](#)]
- Deepa, P.; Gunavathi, C. A systematic review on machine learning and deep learning techniques in cancer survival prediction. *Biophys. Mol. Biol.* **2022**, *174*, 62–71.
- Cox, D. Regression models and life-tables. *J. R. Stat. Soc. Ser. B* **1972**, *34*, 187–202. [[CrossRef](#)]
- Ishwaran, H.; Kogalur, U.B. Random survival forests for R. *R News* **2007**, *7*, 25–31.
- Ishwaran, H.; Kogalur, U.B.; Blackstone, E.H.; Lauer, M.S. Random survival forests. *Ann. Appl. Stat.* **2008**, *2*, 841–860. [[CrossRef](#)]
- Faraggi, D.; Simon, R. A neural network model for survival data. *Stat. Med.* **1995**, *14*, 73–82. [[CrossRef](#)] [[PubMed](#)]
- Katzman, J.L.; Shaham, U.; Cloninger, A.; Bates, J.; Jiang, T.; Kluger, Y. DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **2018**, *18*, 1–12. [[CrossRef](#)] [[PubMed](#)]
- Lee, C.; Zame, W.; Yoon, J.; Van Der Schaar, M. Deephit: A deep learning approach to survival analysis with competing risks. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
- Kim, S.; Song, H.; Kim, S.; Kim, B.; Lee, J.-G. Revisit Prediction by Deep Survival Analysis. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Chengdu, China, 16–19 May 2022; pp. 514–526.
- Wild, C.; Weiderpass, E.; Stewart, B. *World Cancer Report: Cancer Research for Cancer Prevention*; International Agency for Research on Cancer: Lyon, France, 2020.
- Jee, Y.H.; Emberson, J.; Jung, K.J.; Lee, S.J.; Lee, S.; Back, J.H.; Hong, S.; Kimm, H.; Sherliker, P.; Jee, S.H.; et al. Cohort profile: The Korean cancer prevention study-II (KCPS-II) Biobank. *Int. J. Epidemiol.* **2018**, *47*, 385–386f. [[CrossRef](#)]
- He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
- Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)]
- Harrell Jr, F.E.; Lee, K.L.; Califf, R.M.; Pryor, D.B.; Rosati, R.A. Regression modelling strategies for improved prognostic prediction. *Stat. Med.* **1984**, *3*, 143–152. [[CrossRef](#)] [[PubMed](#)]

17. Maalouf, M.; Trafalis, T.B. Rare events and imbalanced datasets: An overview. *Int. J. Data Min. Model. Manag.* **2011**, *3*, 375–388. [CrossRef]
18. SEER\*Explorer SEER Incidence Rates by Age at Diagnosis, 2015–2019. Available online: [https://seer.cancer.gov/statistics-network/explorer/application.html?site=1&data\\_type=1&graph\\_type=3&compareBy=sex&chk\\_sex\\_3=3&chk\\_sex\\_2=2&rate\\_type=2&race=1&advopt\\_precision=1&advopt\\_show\\_ci=on&hdn\\_view=0#graphArea](https://seer.cancer.gov/statistics-network/explorer/application.html?site=1&data_type=1&graph_type=3&compareBy=sex&chk_sex_3=3&chk_sex_2=2&rate_type=2&race=1&advopt_precision=1&advopt_show_ci=on&hdn_view=0#graphArea) (accessed on 2 December 2022).
19. Tabibzadeh, S. Role of autophagy in aging: The good, the bad, and the ugly. *Aging Cell* **2022**, e13753. [CrossRef] [PubMed]
20. Kim, S.-E.; Paik, H.Y.; Yoon, H.; Lee, J.E.; Kim, N.; Sung, M.-K. Sex-and gender-specific disparities in colorectal cancer risk. *World J. Gastroenterol. WJG* **2015**, *21*, 5167. [CrossRef]
21. Suteau, V.; Munier, M.; Briet, C.; Rodien, P. Sex Bias in Differentiated Thyroid Cancer. *Int. J. Mol. Sci.* **2021**, *22*, 12992. [CrossRef]
22. Shobab, L.; Burman, K.D.; Wartofsky, L. Sex Differences in Differentiated Thyroid Cancer. *Thyroid* **2022**, *32*, 224–235. [CrossRef]
23. Rahbari, R.; Zhang, L.; Kebebew, E. Thyroid cancer gender disparity. *Future Oncol.* **2010**, *6*, 1771–1779. [CrossRef]
24. Morganti, S.; Ceda, G.; Saccani, M.; Milli, B.; Ugolotti, D.; Prampolini, R.; Maggio, M.; Valenti, G.; Ceresini, G. Thyroid disease in the elderly: Sex-related differences in clinical expression. *J. Endocrinol. Investig.* **2005**, *28*, 101–104.
25. Yin, D.-T.; He, H.; Yu, K.; Xie, J.; Lei, M.; Ma, R.; Li, H.; Wang, Y.; Liu, Z. The association between thyroid cancer and insulin resistance, metabolic syndrome and its components: A systematic review and meta-analysis. *Int. J. Surg.* **2018**, *57*, 66–75. [CrossRef]
26. Nguyen, D.N.; Kim, J.H.; Kim, M.K. Association of Metabolic Health and Central Obesity with the Risk of Thyroid Cancer: Data from the Korean Genome and Epidemiology Study. *Cancer Epidemiol. Biomark. Prev.* **2022**, *31*, 543–553. [CrossRef]
27. Osorio-Costa, F.; Rocha, G.Z.; Dias, M.M.; Carvalheira, J.B. Epidemiological and molecular mechanisms aspects linking obesity and cancer. *Arq. Bras. Endocrinol. Metabol.* **2009**, *53*, 213–226. [PubMed]
28. Avgerinos, K.I.; Spyrou, N.; Mantzoros, C.S.; Dalamaga, M. Obesity and cancer risk: Emerging biological mechanisms and perspectives. *Metabolism* **2019**, *92*, 121–135. [PubMed]
29. Gallagher, E.J.; LeRoith, D. Obesity and diabetes: The increased risk of cancer and cancer-related mortality. *Physiol. Rev.* **2015**, *95*, 727–748. [PubMed]
30. Moschos, S.J.; Mantzoros, C.S. The role of the IGF system in cancer: From basic to clinical studies and clinical applications. *Oncology* **2002**, *63*, 317–332. [CrossRef]
31. Crosbie, E.J.; Zwahlen, M.; Kitchener, H.C.; Egger, M.; Renehan, A.G. Body Mass Index, Hormone Replacement Therapy, and Endometrial Cancer Risk: A Meta-Analysis. *Cancer Epidemiol. Biomark. Prev.* **2010**, *19*, 3119–3130.
32. Shaw, E.; Farris, M.; McNeil, J.; Friedenreich, C. Obesity and endometrial cancer. *Obes. Cancer* **2016**, 107–136.
33. Liu, Z.; Lin, C.; Suo, C.; Zhao, R.; Jin, L.; Zhang, T.; Chen, X. Metabolic dysfunction-associated fatty liver disease and the risk of 24 specific cancers. *Metabolism* **2022**, *127*, 154955. [CrossRef]
34. He, M.-M.; Fang, Z.; Hang, D.; Wang, F.; Polychronidis, G.; Wang, L.; Lo, C.-H.; Wang, K.; Zhong, R.; Knudsen, M.D.; et al. Circulating liver function markers and colorectal cancer risk: A prospective cohort study in the UK Biobank. *Int. J. Cancer* **2021**, *148*, 1867–1878. [CrossRef]
35. Stocker, R.; Yamamoto, Y.; McDonagh, A.F.; Glazer, A.N.; Ames, B.N. Bilirubin is an antioxidant of possible physiological importance. *Science* **1987**, *235*, 1043–1046. [CrossRef]
36. Horsfall, L.J.; Rait, G.; Walters, K.; Swallow, D.M.; Pereira, S.P.; Nazareth, I.; Petersen, I. Serum bilirubin and risk of respiratory disease and death. *JAMA* **2011**, *305*, 691–697. [CrossRef]
37. Sarna, L.; Evangelista, L.; Tashkin, D.; Padilla, G.; Holmes, C.; Brecht, M.L.; Grannis, F. Impact of respiratory symptoms and pulmonary function on quality of life of long-term survivors of non-small cell lung cancer. *Chest* **2004**, *125*, 439–445. [CrossRef] [PubMed]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.