# Identification of mortality-risk-related missense variant for renal clear cell carcinoma using deep learning

Ther Adv Chronic Dis

2021, Vol. 12: 1–14

DOI: 10.1177/2040622321992624

Jin-Bor Chen, Huai-Shuo Yang, Sin-Hua Moi, Li-Yeh Chuang and Cheng-Hong Yang

*Abstract*

**Introduction:** Kidney renal clear cell carcinoma (KIRCC) is a highly heterogeneous and lethal cancer that can arise in patients with renal disease. DeepSurv combines a deep feed-forward neural network with a Cox proportional hazards function and could provide optimized survival results compared with convenient survival analysis.

**Methods:** This study used an improved DeepSurv algorithm to identify the candidate genes to be targeted for treatment on the basis of the overall mortality status of KIRCC subjects. All the somatic mutation missense variants of KIRCC subjects were abstracted from TCGA-KIRC database.

**Results:** The improved DeepSurv model (95.1%) achieved greater balanced accuracy compared with the DeepSurv model (75%), and identified 610 high-risk variants associated with overall mortality. The results of gene differential expression analysis also indicated nine KIRCC mortality-risk-related pathways, namely the tRNA charging pathway, the D-myo-inositol-5-phosphate metabolism pathway, the DNA double-strand break repair by nonhomologous end-joining pathway, the superpathway of inositol phosphate compounds, the 3-phosphoinositide degradation pathway, the production of nitric oxide and reactive oxygen species in macrophages pathway, the synaptic long-term depression pathway, the sperm motility pathway, and the role of *JAK2* in hormone-like cytokine signaling pathway. The biological findings in this study indicate the KIRCC mortality-risk-related pathways were more likely to be associated with cancer cell growth, cancer cell differentiation, and immune response inhibition.

**Conclusion:** The results proved that the improved DeepSurv model effectively classified mortality-related high-risk variants and identified the candidate genes. In the context of KIRCC overall mortality, the proposed model effectively recognized mortality-related high-risk variants for KIRCC.

***Keywords:*** Kidney renal clear cell carcinoma, deep learning, survival analysis

Received: 20 April 2020; revised manuscript accepted: 13 January 2021.

Correspondence to:
**Cheng-Hong Yang**
Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, 415 Jiangong Road, San-Min District, Kaohsiung, 82444

Biomedical Engineering, Kaohsiung Medical University, Kaohsiung

Drug Development and Value Creation Research Center and with PhD Program in Biomedical Engineering, Kaohsiung Medical University, Kaohsiung 80341, Taiwan
**chyang@nkust.edu.tw**

**Jin-Bor Chen**
Division of Nephrology, Department of Internal Medicine, Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Kaohsiung

**Huai-Shuo Yang**
Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung

**Sin-Hua Moi**
**Li-Yeh Chuang**
Department of Chemical Engineering and Institute of Biotechnology and Chemical Engineering, I-Shou University, Kaohsiung

## Introduction

Kidney renal clear cell carcinoma (KIRCC) is a type of lethal genitourinary disease and is the leading cause of malignant kidney tumors. Published studies have indicated that KIRCC recognition could be increased by identifying inter- and intra-tumor molecular heterogeneity.[1,2] If KIRCC is diagnosed at an early stage, surgery may effectively eliminate cancer from the patient's body. However, the rate at which cancer can be eliminated becomes worse in later stages, and fewer than 20% of patients with metastatic KIRCC have a survival time longer than 2 years.[3,4]

The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma (TCGA-KIRC) project has assembled large-scale sequencing data containing multiple data types—for instance, data concerning DNA methylation, clinical information, and other forms of genomic information; this data set enables the discovery of new molecular mechanisms of KIRCC.[5] One study indicated KIRCC is an immune-responsive disease and can potentially be treated using immune inhibitors.[6] Furthermore, hemodialysis has been widely researched in many studies,[7–11] and the KIRC database has been extensively used for the comprehensive molecular characterization of KIRCC.[12,13] Hence, the genomic characteristics and molecular pathways of KIRCC, especially the immune-checkpoint-related genes, should be further investigated.

It is difficult to efficiently apply conventional analytic approaches to high-throughput and high-variability genomic data; however, machine learning is practical for this purpose.[14] Machine learning can extract complex features from high-throughput genomic data[15,16] and has been widely used in genomics research.[17,18] The deep learning (DL) algorithm is one of the most useful machine learning approaches in genomic studies.[19,20]

Accurate identification of mortality-related missense variants is a primary objective for evaluating the result of a specific disease.[21] In cancer studies, the outcome under assessment is mainly concerned for the time to some specific event of interest, such as mortality.[22,23] Time-to-events models for evaluating survival analysis have been extensively used to produce reliability models in biomedicine.[24–26] In survival analysis, log-rank tests, Kaplan–Meier plots, Cox models, and survival tree analysis[27,28] are commonly used methods for estimating time-to-events data.[29] The most widely employed method in this context is the semiparametric Cox proportional hazards regression (CoxPH) model,[30] which is employed to estimate the time-varying effects of observed features on the risk of an occurred event. Most CoxPH model applications lack hazard proportionality and ignore interactions between risk features; these deficiencies may increase the possibility of incorrect assessment of mortality risk with assumptions of linearity. Therefore, nonlinear log-risk functions are required to accurately fit survival data to improve the performance of survival models.[31,32]

Researchers have developed nonlinear survival models with neural networks such as the Faraggi–Simon Network[33–36] and deep neural networks. DL[37] was developed from neural networks and provides favorable outcome estimation in survival analysis. DeepSurv, an extension of DL-based survival analysis[31] that combines a CoxPH model with a modern DL algorithm, has been used to estimate the survival risks with a recommender system. DeepSurv predicted outcomes accurately by applying both linear and nonlinear survival analysis methods to survival data.[31] However, in DeepSurv, an "internal covariate shift" problem may occur because of variation in the input distributions of each layer during the training procedure; this might render the model training procedure slow and unstable.[38]

The development of machine learning techniques has allowed modeling of various intricate nonlinear relationships. Machine learning methods have enhanced the overall prediction quality for many practical applications in diverse domains. In common applications, such as classification and regression, machine learning is effective when given a sufficiently large set of training instances in a reasonable dimensional feature space. However, in survival analysis, the machine learning methods inevitably face the additional challenge of dealing with censored instances and model time estimation.[39] DL has been applied in survival analysis. Numerous methods have been proposed, such as SurvivalNet,[40] DeepHit,[41] and DeepSurv.[31] DeepSurv was inspired by Faraggi–Simon networks. Both DeepSurv and Faraggi–Simon networks require training of the network and combining the network with a CoxPH model, whereas DeepSurv improves the model with modern DL techniques. This study applied an improved DL-based survival analysis to identify mortality-risk-related missense mutation variants and determine the differential expression of candidate genes from TCGA-KIRC.

## Results

### *TCGA-KIRC data set*

The Cancer Genome Atlas data portal is an open access platform, and all data sets are available for download at https://tcga-data.nci.nih.gov/tcga/. The comprehensive molecular characterization of KIRCC is described, and the detailed information can be reviewed at https://tcga-data.nci.nih.gov/tcga/tcga DataType.jsp. The publicly available KIRCC data set in the TCGA database was used as the major data source for this study. In the relevant TCGA-KIRC data set, all detected missense mutation variants from the DNA-seq data set were analyzed and had accepted proper treatments based on the medical treatment guidelines for cancer. DNA-seq expression refers to genomic data obtained from the DNA methylation (Illumina Human Methylation 450) pipeline in the TCGA database. We selected the following data sets to represent DNA-seq expression for our analysis: biotype, mutation calling 3 (MC3) overlap, PICK, scale-invariant feature transform (SIFT) score, polymorphism phenotyping (PolyPhen) score, and mutation score. Regarding clinical characteristics, all discovered missense mutation variants DNA-seq genomic data in the kidney cancer subjects were acquired from the TCGA database and paired with each other using the defined barcode of each data set.

The final features sets included gender (i.e. male and female), race (Asian, white, and black or African-American), tumor stage (according to the American Joint Committee on Cancer staging), biotype (containing protein coding, polymorphic pseudogene, nonsense-mediated decay, IG C gene, IG V gene, TR C gene and TR V gene), MC3 overlap (indicative of whether the specified region was overlapped with a multicenter-mutation-calling variant for the same sample pair), PICK (which explains whether a particular block of consequent data had been selected by the picked feature of the variant effect predictor), age group (younger: subjects aged less than 50 years; elder: subjects aged more than 50 years), SIFT score, PolyPhen score, and mutation score. The follow-up intervals of all subjects with kidney cancer were such that they were tracked from the initial diagnosis date to the date of death or to the end of the study. Subjects lost to follow-up before the end of the study were regarded as right-censored subjects.

In this study, we transformed our TCGA-KIRC data set into two forms, binary and mixed-type. In our binary TCGA-KIRC data set, all features were dichotomous, determined on the basis of subgroup similarity of categorial features or the optimal cutoff of the enrolled subjects; and in our mixed-type TCGA-KIRC data set, we retained the original features to retain diversity.

### *Feature set and outcome distribution in TCGA-KIRC*

The distribution of the clinical features and, DNA-seq expression of TCGA-KIRC missense mutation variants according to cancer mortality status are summarized in Table 1. The results indicated that the living and deceased subjects were significantly different in terms of the distributions of gender, race, tumor stage, MC3 overlap, PICK, age group, and mutation score. According to the results, male characteristics represented significantly higher proportions in mortality-related variants compared with female characteristics; white racial features represented higher proportions of risk-related variants than black or African-American did. Asian feature did not obtain any risk-related variants in this data set; the clear cell adenocarcinoma characteristics retained some extremely significant risk-associated variants; in terms of tumor stage features, stage I and stage IV obtained a highly significant death-related variants with proportion of 39.96% and 41.04%, respectively; the elder subject characteristics obtained the greatest proportion of mortality-associated missense variants.

### *Performance comparison between survival models in TCGA-KIRC*

A comparison of the four survival models' performance levels is presented in Table 2 and Figure 1 As shown in Table 2, the DeepSurv binary input model obtained a confusion matrix [true positive (TP) = 29, false positive (FP) = 27, false negative (FN) = 64 and true negative (TN) = 1421] and a C-index of 77.5%; the improved DeepSurv binary input model obtained a confusion matrix (TP = 27, FP = 26, FN = 66, and TN = 1422) and a C-index of 77.5%; the DeepSurv mixed-type input model obtained a confusion matrix (TP = 47, FP = 8, FN = 46, and TN = 1440) and a C-index of 93.1%; and the improved DeepSurv mixed-type model obtained a confusion matrix (TP = 86, FP = 33, FN = 7, and TN = 1415) and a C-index of 98.7%.

Figure 1(a) shows the normalized confusion matrix heatmap of the four survival models. As shown in Figure 1(b) and (c), the DeepSurv binary input model obtained a balanced accuracy of 64.7%, a balanced error rate of 35.3%, a sensitivity of 31.2%, and a specificity of 98.1%; the improved DeepSurv binary input model obtained a balanced accuracy of 63.6%, a balanced error rate of 36.4%, a sensitivity of 29%, and a specificity of 98.2%; the DeepSurv mixed-type input model obtained a balanced accuracy of 75%, a balanced error rate of 25%, a sensitivity of 50.5%, and a specificity of 99.5%; and the improved DeepSurv mixed-type input model obtained a

**Table 1.** Baseline clinical characteristics and DNA-seq mutation score of kidney cancer missense mutation variants according to The Cancer Genome Atlas Kidney Renal Clear Cell Carcinoma (TCGA-KIRC) cancer mortality status.

| Features | Category | Alive (*n* = 7241) | Dead (*n* = 463) | *p*-Value |
|---|---|---|---|---|
| Gender | Male | 4783 (66.05%) | 419 (90.5%) | **<0.001**[a] |
| | Female | 2458 (33.95%) | 44 (9.5%) | |
| Race | Asian | 152 (2.1%) | – | **0.003**[b] |
| | Others | 7089 (97.9%) | 463 (100%) | |
| Race (mixed-type) | Asian | 152 (2.1%) | – | **<0.001**[b] |
| | White | 6262 (86.48%) | 293 (63.28%) | |
| | Black or African-American | 827 (11.42%) | 170 (36.72%) | |
| Tumor stage | Stage I–III | 6900 (95.29%) | 273 (58.96%) | **<0.001**[a] |
| | Stage IV | 341 (4.71%) | 190 (41.04%) | |
| Tumor stage (mixed-type) | Stage I | 4518 (62.39%) | 185 (39.96%) | **<0.001**[b] |
| | Stage II | 802 (11.08%) | 88 (19%) | |
| | Stage III | 1580 (21.82%) | – | |
| | Stage IV | 341 (4.71%) | 190 (41.04%) | |
| Biotype | Protein coding | 7197 (99.39%) | 462 (99.78%) | 0.449[b] |
| | Others | 44 (0.61%) | 1 (0.22%) | |
| Biotype (mixed-type) | Protein coding | 7197 (99.39%) | 462 (99.78%) | 0.772[b] |
| | Polymorphic pseudogene | 1 (0.01%) | – | |
| | Nonsense-mediated decay | 17 (0.24%) | – | |
| | IG C gene | 7 (0.09%) | – | |
| | IG V gene | 12 (0.18%) | – | |
| | TR C gene | 1 (0.01%) | – | |
| | TR V gene | 6 (0.08%) | 1 (0.22%) | |
| MC3 Overlap | No | 252 (3.48%) | 11 (2.38%) | 0.256[a] |
| | Yes | 6989 (96.52%) | 452 (97.62%) | |
| PICK | No | 1753 (24.21%) | 93 (20.09%) | 0.054[a] |
| | Yes | 5488 (75.79%) | 370 (79.91%) | |
| Age group | Younger | 1195 (16.5%) | 30 (6.48%) | **<0.001**[a] |
| | Elder | 6046 (83.5%) | 433 (93.52%) | |
| Age | mean ± std | 60.45 ± 10.84 | 66.98 ± 13.78 | **<0.001**[c] |

*(Continued)*

**Table 1.** (Continued)

| Features | Category | Alive ($n=7241$) | Dead ($n=463$) | $p$-Value |
|---|---|---|---|---|
| Age normalization | mean ± std | 3.30e-16 ± 1 | 4.14e-16 ± 1 | |
| SIFT | Low | 3925 (54.21%) | 236 (50.97%) | 0.192[a] |
| | High | 3316 (45.79%) | 227 (49.03%) | |
| SIFT (mixed-type) | mean ± std | 0.14 ± 0.24 | 0.17 ± 0.25 | 0.071[c] |
| PolyPhen | Low | 3531 (48.76%) | 238 (51.4%) | 0.292[a] |
| | High | 3710 (51.24%) | 225 (48.6%) | |
| PolyPhen (mixed-type) | mean ± std | 0.53 ± 0.42 | 0.5 ± 0.42 | 0.126[c] |
| Mutation score | Low | 4015 (55.45%) | 253 (54.64%) | 0.772[a] |
| | High | 3226 (44.55%) | 210 (45.36%) | |
| Mutation score (mixed-type) | mean ± std | 0.21 ± 0.19 | 0.21 ± 0.17 | 0.638[c] |

*p*-Value is estimated using [a]chi-squared, [b]fisher's exact, or [c]independent two-sampled *t*-test appropriately, bold indicates the significant difference.

**Table 2.** Comparison of performance of TCGA-KIRC classification models based on DeepSurv.

| Classification model | TP | FP | FN | TN | C-index (%) |
|---|---|---|---|---|---|
| Binary | | | | | |
| DeepSurv | 29 | 27 | 64 | 1421 | 77.5 |
| Improved DeepSurv | 27 | 26 | 66 | 1422 | 77.5 |
| Mixed type | | | | | |
| DeepSurv | 47 | 8 | 46 | 1440 | 93.1 |
| Improved DeepSurv | 86 | 33 | 7 | 1415 | 98.7 |

FN, false negative; FP, false positive; TN, true negative; TP, true positive.

balanced accuracy of 95.1%, a balanced error rate of 4.9%, a sensitivity of 92.5%, and a specificity of 97.7%. Our improved DeepSurv mixed-type input model obtained the overall best performance of the four survival models.

### *Performance comparison between risk models in TCGA-KIRC*

As shown in Figure 2, the comparison of cancer mortality between high-risk and low-risk categories was made using a Kaplan–Meier curve and a log-rank test. All the risk models exhibited significantly lower survival rates (indicating high mortality rates) in the high-risk category than in the low-risk category. The improved DeepSurv model with the mixed-type data set obtained the best performance of the four risk models.

### *Performance comparison between risk models in TCGA-KIRC*

According the distinguish results of the mixed-type data set based on improved DeepSurv, the genes for which high-risk missense mutation variants overlapped in all classification models were

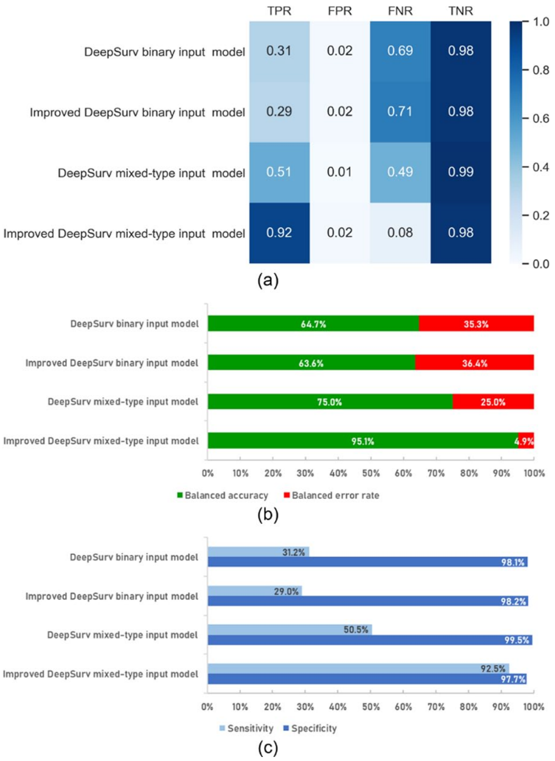

**Figure 1.** (a) Heatmap of the normalized confusion matrix in comparison of TCGA-KIRC classification models based on DeepSurv. (b) Stacked bar chart of the balanced accuracy and balanced error rate in comparison of TCGA-KIRC classification models based on DeepSurv. (c) Bar chart comparing the specificity and sensitivity of TCGA-KIRC classification models based on DeepSurv.

FNR, false negative rate; FPR, false positive rate; TNR, true negative rate; TPR, true negative rate.

selected as the candidate genes ($n=580$) for mortality risk estimation in TCGA-KIRC. The differential expression analysis between tumor and normal tissue was conducted for the candidate genes to further understand the gene function. The improved DeepSurv model identified 610 high-risk variants according to the overall mortality of TCGA-KIRC subjects. The results of gene differential expression analysis indicated nine KIRCC mortality-risk-related pathways, namely the tRNA charging pathway, the D-myo-inositol-5-phosphate metabolism pathway, the DNA double-strand break repair by nonhomologous end-joining pathway, the superpathway of inositol phosphate compounds, the 3-phosphoinositide degradation pathway, the production of nitric oxide and reactive oxygen species in macrophages pathway, the synaptic long-term depression pathway, the sperm motility pathway, and the role of JAK2 in hormone-like cytokine signaling pathway. The biological findings in this study indicate the KIRCC mortality-risk-related pathways were more likely to be associated with cancer cell growth, cancer cell differentiation, and immune response inhibition. The detail of the gene ontology (GO) and gene set enrichment analysis (GSEA) are presented in Supplemental Table S1.

## Discussion

This study applied DeepSurv and the proposed improved DeepSurv algorithms to identify high-risk missense mutation variants and candidate genes in mortality risk. In our data preprocessing, we transformed the data set into two types: binary and mixed-type. Although the clear distribution of features and outcomes could be given by using the dichotomous procedure in the binary data set, the mixed-type data set retained its diversity of features and contributed to training the desirable models. In DeepSurv, the deep neural network learned the nonlinear weights and biases and then estimated the log-risk function through the Cox proportional hazards function. It was proved that DeepSurv could provide the same or even better outcome performance than previous linear or nonlinear survival algorithms.[31] As a baseline survival model, DeepSurv demonstrated its generalization ability. Relatedly, BatchNorm is an efficient learning technique widely used in training models. It can accomplish numerous advantageous functions, such as training the network rapidly, enabling a high learning rate, facilitating weight initialization, making numerous activation functions viable, simplifying the creation of deep networks, providing regularization, and eliminating the necessity of dropout.[42] In the improved DeepSurv, we took the advantage of DeepSurv and imported BatchNorm techniques for model training and obtained excellent outcomes. As the analysis results proved, mixed-type input models performed much better than binary input models; the improved DeepSurv model was superior to the original DeepSurv model. Due to the dichotomous procedure of the binary data set, the reduced diversity probably eliminated some information concerning clinical features. Although the balanced accuracy of the improved DeepSurv model was 1.1% worse than that of DeepSurv, owing to some dichotomous information missing from the binary data set, the balanced accuracy of the improved DeepSurv was 20.1% better than

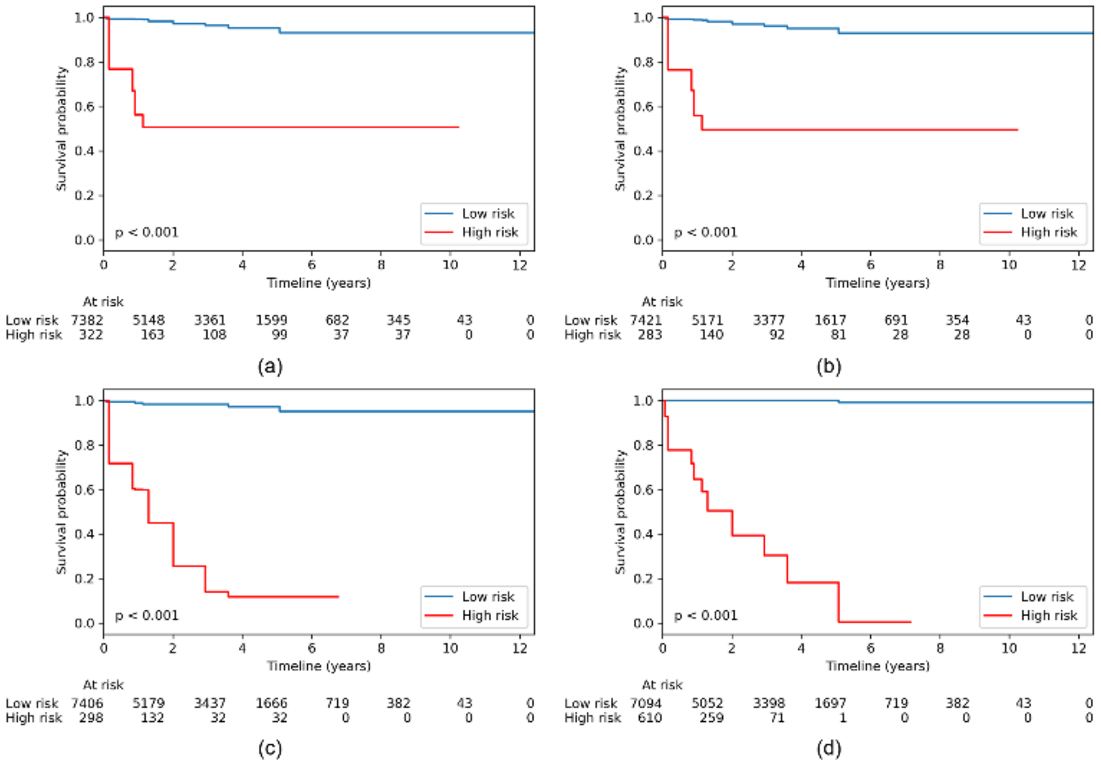

**Figure 2.** (a) Kaplan–Meier curve of TCGA-KIRC based on the DeepSurv binary input model. (b) Kaplan–Meier curve of TCGA-KIRC based on the improved DeepSurv binary input model. (c) Kaplan–Meier curve of TCGA-KIRC based on the DeepSurv mixed-type input model. (d) Kaplan–Meier curve of TCGA-KIRC based on the improved DeepSurv mixed-type input model.

that of the original DeepSurv on the mixed-type data set.

The model indicated tRNA charging, D-myo-inositol-5-phosphate metabolism, DNA double-strand break repair by nonhomologous end joining, the superpathway of inositol phosphate compounds, 3-phosphoinositide degradation, the production of nitric oxide and reactive oxygen species in macrophages, synaptic long-term depression, sperm motility, and the role of JAK2 in hormone-like cytokine signaling pathways might relate to KIRCC mortality risk. Some studies have indicated that tRNA charging participates in tumorigenesis processes and can regulate oncogenic mutations by playing crucial roles in suppressing proliferation and growth when intracellular supplies of essential metabolites become reduced.[43,44] D-myo-inositol-5-phosphate metabolism was enriched in differentially expressed genes of insulin molecules.[45] DNA double-strand breaks are the most deleterious DNA lesions; they can lead to genomic instability and carcinogenesis. Nonhomologous end joining is the major repair pathway in mammalian cells; it can be induced by endogenous and exogenous agents.[46] Therefore, the DNA double-strand break repair by nonhomologous end-joining pathway was considered to play roles in KIRC mortality risk regulation. Both the superpathway of inositol phosphate compounds and 3-phosphoinositide degradation were enriched in distinct skeletogenesis pathways.[47] The production of nitric oxide and reactive oxygen species in macrophages was associated with the NADPH oxidase 2 pathway in renal oxidative stress in *Aqp11*$^{-/-}$ mice.[48] Synaptic long-term depression was associated with adipose tissue DNA methylome changes in the development of diabetes.[49] Sperm motility was proved to have a significant relationship with kidney transplantation.[50] JAK2 is a set of nonreceptor protein tyrosine kinases from the Janus kinase (JAK) family, and this set was reported to play a role in hormone-like cytokine signaling associated with SOX2-regulated transcriptome in glioma stem cells.[51] Moreover, the JAK/signal transducer and activator of transcription (STAT) signaling pathway is also involved in cell growth, cell differentiation, and immune functions.[52,53] All the identified biological pathways derived from the

KIRCC mortality-risk-related candidate gene set were directly or indirectly associated with cancer cell growth, invasion, and immune function. Hence, the participating genes in the identified pathways might have novel potential in anticancer research for KIRCC.

The present study must acknowledge several limitations. The induction and development of KIRCC are associated with multiple genetic variations that are combined with environmental risk factors and behaviors (including chronic inflammation) and play roles in the activation of oncogenes or tumor suppressor genes. Because the current study was retrospective, the study might have ignored some confounding environmental factors that had not been recorded in the data sets. An imbalanced data set could lead to statistically imbalanced results in terms of sensitivity and specificity. Differentially censored subjects or subjects lost to follow-up could also bias the study results. In addition, the utility of risk and classification models might require additional experimental and clinical proof. Our study was limited by its retrospective analysis and unavailability of clinical parameters. Hence, relevant factors including MSKCC, IMDC, and Karnofsky scores could not be analyzed in our study. We believe that these factors contribute to mortality risk in renal clear cell carcinoma. However, a previous study demonstrated that different models can yield dissimilar prognoses on the basis of the inclusion of different clinical parameters.[54] Accordingly, our study focused on the missense variants of candidate genes. Further research using the aforementioned models with missense variants is warranted to examine survival prognosis in KIRCC. Despite the aforementioned limitations, this study generated an improved DeepSurv algorithm for identifying high-risk missense mutation variants and candidate genes using genomic data.

In cancer medicine, the primary challenge for realizing the genetic basis of carcinoma and making new breakthroughs is the application of next-generation sequencing data. In the current study, we proposed our improved DeepSurv algorithm for effectively identifying missense mutation variants related to cancer mortality and immunologic signatures with genomic data from TCGA-KIRC. New targets for anticancer treatment using immunologic or antiangiogenic mechanisms were provided by the identified canonical pathways identified by the improved DeepSurv. Further studies are required to interpret the interactions between the identified pathways and the innate immune system to improve the distinguishability of potential variants and make new breakthroughs in anticancer therapy. Future studies should enhance the improvement of survival model performance. We can focus on various aspects of model training, such as applying grid search optimization to systematically tune various hyperparameters, such as model architecture, activation function, learning rate, batch size, and optimizer.

Furthermore, in fact, DeepSurv is constrained by the proportional assumption of its CoxPH model, whereas some other studies have extended CoxPH models to eliminate the proportional restriction.[32] The results of the present study suggest that loss function research might be advanced by the combination of DL and survival models. The successful analysis of genomic data depends on accurate and efficient algorithms; the proposed algorithms should achieve comprehensive estimation based on genomic data. In future studies, the proposed algorithms must precisely identify risk-related variants of KIRCC mortality.

This study proposed an improved DeepSurv model to identify high-risk missense mutation variants for overall mortality of KIRCC. The performance of the DeepSurv model and the improved DeepSurv model were compared by analyzing two types of data sets. The results indicated that the models applied to a mixed-type data set could be trained better than the models applied to a binary data set due to more detailed features in the mixed-type data set. In addition, the improved DeepSurv model exhibited a superior classification ability for mortality-related high-risk variants and candidate gene identification. The biological findings in this study indicate the KIRCC mortality-risk-related pathways were more likely to be associated with cancer cell growth, differentiation, and immune response inhibition. Thus, the KIRCC candidate genes related to mortality risk determined by the improved DeepSurv model might provide novel targets for further research. In conclusion, the proposed model is beneficial for the recognition of mortality-related high-risk variants for the overall mortality of KIRCC and precise identification of KIRCC variants related to mortality risk.

## Methods

### *Data preprocessing*

The distribution of missense mutation variant features was summarized by frequency and percentage according to their vital status. The difference between categories was estimated using Pearson's chi-squared test. The performance of the risk models was determined using an accuracy test, where the risk models with high accuracy were considered likely to classify high-risk and low-risk mutation variants and candidate genes accurately. Candidate genes were defined as those that were recognized as belonging to the high-risk category in all risk models. GO and GSEA were conducted using the candidate gene set to further explore some pathways potentially related to cancer mortality. All the analyses were performed using PyTorch (version. 1.3),[55] TCGAbiolinks and the related packages in the R software environment (version. 3.5.3). In TCGA-KIRC data preprocessing, all data sets were transformed into two forms (binary and mixed-type). Preprocessing also normalized the transformed nominal-to-numerical features into values ranging from 0 to 1. In binary data sets, all features were dichotomous according to the subgroup similarity of categorial features or the optimal cutoff of the enrolled subjects. The mixed-type data set retained the original normalized numerical features. The distribution of features between the alive and dead groups was estimated using a chi-square test, Fisher's exact test, or an independent two-sample *t*-test. The follow-up intervals of all subjects with kidney cancer were tracked from the initial diagnosis date to the death date or the end of the study.

### *Survival analysis*

In survival analysis (time-to-event analysis), survival data are composed of three major elements: (1) an individual's baseline data $x$, which describes the relationships of survival distributions to features; (2) a failure event time $T$, which records the time elapsed between the time from data collection to the event occurrence or the latest diagnosis date, and (3) an event indicator $E$, which denotes whether the event (e.g. death) is observed or not.

Survival and hazard functions are the two primary functions in survival analysis. The survival function is defined as $S(t) = \Pr\ (T > t)$ which denotes to the probability that an individual survives longer than the time $t$. The hazard function $\lambda(t)$ denotes the instantaneous probability that the event occurs at time $t$ but has not occurred before time $t$, defined as follows:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t T \geq t)}{\Delta t} \qquad (1)$$

where $t$ is the time that an individual has already survived and $\Delta t$ is an extra infinitesimal amount of time. The hazard function estimates the probability of mortality; a high hazard indicates a higher risk of mortality.

### *Survival models*

In survival models, proportional hazards models are usually employed to model the hazard function. A typical proportional hazards model supposes the hazard function consist of two units: (1) the baseline hazard function $\lambda_0(t)$, describing how the risk of event per time unit changes over time at baseline levels of features, and (2) the risk score $r(x) = e^{\mathrm{h}(x)}$ which $h(x)$ is the log-risk function that describes the effect of an individual's features on the baseline hazard. The hazard function is defined as the follows:

$$\lambda(t|x) = \lambda_0(t) \cdot e^{h(x)} \qquad (2)$$

Survival models can be divided into linear and nonlinear types. In linear survival models, the Cox proportional hazards regression (CoxPH) model is a semiparametric approach[30] that commonly uses a linear function $\hat{h}_\beta(x) = \beta^{\mathrm{T}} x$ to estimate the log-risk function $h(x)$

$$\begin{aligned} L_{Cox}(\beta) &= \prod_{i:E_i=1} \frac{\hat{r}_\beta(x_i)}{\sum_{j \in \Re(T_i)} \hat{r}_\beta(x_j)} \\ &= \prod_{i:E_i=1} \frac{e^{\hat{h}_\beta(x_i)}}{\sum_{j \in \Re(T_i)} e^{\hat{h}_\beta(x_j)}} \end{aligned} \qquad (3)$$

where $x_i$, $T_i$, and $E_i$, respectively, signify the baseline data, event time and event indicator in the $i$-th observation. The product is measured in the set of individuals with the observable event $E_i = 1$. The risk set $\Re(t) = \{i \mid T_i \geq t\}$ represents the set of individuals still at risk of mortality at time $t$. However, because most applications are nonlinear, using a linear proportional hazards model to model nonlinear gene interaction, for example, may not be appropriate. In nonlinear survival models, the Faraggi–Simon method first combines a neural network with a CoxPH function. Hence, nonlinear output can be generated to construct a nonlinear proportional hazard model. Scholarly papers have argued that Faraggi–Simon networks do not exhibit superior performance to the linear CoxPH.

### DeepSurv

DeepSurv is a deep feed-forward neural network combined with a Cox proportional hazards function.[31] The network architecture is similar to that of the Faraggi–Simon method, but DeepSurv can be constructed with more than one hidden layer and can exploit the novel DL techniques. The output of the network is a single neuron, which estimates the log-risk function $\hat{h}_\theta(x)$ in the hazards function (2). The network is trained and optimized by setting the loss function as the average negative log version of the Cox partial likelihood (3) and with an additional $l_2$ regularization as follows:

$$l(\theta):=-\frac{1}{N_{E=1}}\sum_{i:E_i=1}\left(\begin{array}{l}\hat{h}_\theta(x_i)\\ -log\sum\limits_{j\in\Re(T_i)} e^{\hat{h}_\theta(x_j)}\end{array}\right) + \lambda\cdot\|\theta\|_2^2 \tag{4}$$

where $N_{E=1}$ denotes the number of individuals with an observable event and $\lambda$ is the $l_2$ regularization parameter. The weights of DeepSurv can be trained and optimized by minimizing the output loss (4) using optimization algorithms.

### Improved DeepSurv

For our improved DeepSurv, we enhanced the baseline DeepSurv by adding a batch normalization (BatchNorm) layer.[56] In DL training, internal covariate shift usually occurs because of the distribution of each layer's input changes. BatchNorm is an extensively used technique in deep neural network training. It can address the internal covariate shift problem by normalizing layer input and enables the use of much higher-than-typical learning rates and the performance of initialization with less than usual carefulness. In our improved DeepSurv model, each BatchNorm layer was added before each activation function to prevent gradients from vanishing or exploding.

### Mortality risk recommender system

To identify the missense variants, we developed a mortality risk recommender system to classify the individual vital status according to the predicted individual survival rate at the final observed time point $S(t_{max}|x)$. The recommender system function $\text{rec}_{class}$ can be described as follows:

$$\text{rec}_{\text{class}}(\text{x})=\begin{cases}\text{High risk, if S}(\text{t}_{max}|\text{x})\le 0.5\\ \text{Low risk, if S}(\text{t}_{max}|\text{x})>0.5\end{cases} \tag{5}$$

Hence, we can use the obtained $\text{rec}_{class}$ to identify whether the missense variants are at high risk. If the predicted survival rate is more than 0.5, we classify the vital status as low risk. If the predicted survival rate is less than 0.5, we classify the vital status as high risk.

### Model architecture and hyperparameter configuration

In this study, we employed a baseline DeepSurv and an improved DeepSurv. The same hyperparameters were configured in both baseline DeepSurv and improved DeepSurv models, but the baseline DeepSurv was trained without using the BatchNorm technique. In the deep neural network architecture, we constructed each input layer with ten neurons for the ten features, the four hidden layers with eight neurons, and each network's single output neuron for log-risk estimation. In both models, a rectified linear unit (ReLU)) function was behind each hidden fully connected layer; for the improved DeepSurv, BatchNorm was additionally inserted before each ReLU layer. In the training process, survival models were trained with the following hyperparameters: the Adam optimizer[57] was configured with a learning rate of 0.001, the batch size of 512 and 10,000 epochs in the training and validation sets. The procedure for training models and the subsequently yielded predictions are described in Algorithm 1.

### Model performance evaluation

Kaplan–Meier estimator is a commonly used nonparametric statistical method to measure the survival function from survival data.[58] They described the term "death" as a metaphor for any potential event that might be subject to random sampling, especially when all individuals of the random sample could not be entirely observed. Incomplete observation usually occurs because

**Algorithm 1.** Improved DeepSurv algorithm.

```
Input: an individual baseline data $x$, a failure event time $T$, an event indicator $E$.
Output: a single node $\hat{h}_\theta(x)$
Divide TCGA-KIRC into binary or mixed-type dataset
Divide dataset into trainset, validset and testset
Define model ← DeepSurv or improved DeepSurv
Define loss function ← CoxPH function
# Train and save the best performance model
for epoch = 1 → epochs do
    # Training Phase
    foreach $x_{batch}$, ($T_{batch}$, $E_{batch}$) in trainset do
        output = model ($x_{batch}$)
        loss = loss function (output, ($T_{batch}$, $E_{batch}$))
        back-propagation()
    end foreach
    # Validation Phase
    foreach $x_{batch}$, ($T_{batch}$, $E_{batch}$) in validset do
        output = model ($x_{batch}$)
        loss = loss function (output, ($T_{batch}$, $E_{batch}$))
    end foreach
end for
# Prediction
foreach $x$ in testset do
    survival rate = model.predict(x)
    if survival rate$_{(Tmax)}$ < 0.5 do
        predict result ← $E$ = 1
    end if
    else do
        predict result ← $E$ = 0
    end else
end foreach
```

the contact with some sample individuals has been lost before the event, other intervention variables affect the event, or insufficient data result from observing the event in all sample individuals in a given length of time. Medical researchers evaluate the influence of an intervention by estimating the number of individuals that survived after that intervention over a period. The Kaplan–Meier survival curve represents the probability of surviving for some particular duration while considering time as many small intervals. The Kaplan–Meier estimator was mainly used to evaluate the statistical significance of results in this survival analysis research.

The C-statistic[59] (also known as the concordance statistic or C-index) is the most frequently used evaluation metric to assess the discriminatory power of a logistic regression predictive model in survival analysis. In medical research, with C-statistic, a randomly selected individual who underwent an event is assigned a higher risk score than an individual who did not undergo the event. The C-statistic can consider censored data and is generally regarded as the area under the receiver operating characteristic curve within a Cox model. The value of the C-static estimation is described in the following passages. A value lower than 0.5 signifies an especially poor model. A value of 0.5 indicates that the model predicts the outcome with accuracy close to that of random choice. A value over 0.7 indicates a useful model. A value over 0.8 indicates a strong model. A value of 1 indicates a perfectly predictive model. We used the C-statistic to evaluate the performance of the models.

### *Gene ontology and pathway annotation for candidate genes*

GO's gene annotation classification provides a set of tools that can be used to systematically analyze gene functions.[60] The attributes of each gene are stored in a tree-like database in a meticulously structured manner. In this experiment, we used the selected candidate genes to perform GO. GSEA is a powerful analytical approach for interpreting gene expression data.[61] This approach focuses on gene sets (i.e. genomes with common biological functions, chromosomal positions, or regulatory roles). GSEA offers insight into numerous cancer-related data sets, whereas single-gene analysis has found little similarity between any two independent studies on the survival rate of cancer patients. GSEA determines whether the genes in each gene set are enriched in the upper or lower part of the gene list after the phenotypic relevance ranking, the effect of the cooperative changes of genes in the gene set on the phenotypic change are then judged. In this study, GO and GSEA was employed to reveal gene ontology annotation and biological pathways.

## Author contributions

J.-B.C., L.-Y.C. and C.-H.Y. developed the study concept and design; performed experiments; and drafted the manuscript. H.-S.Y. and S.-H.M. analyzed and interpreted the data. All authors read and approved the final manuscript.

## Conflict of interest statement

The authors declare that there is no conflict of interest.

## Ethics statement

The Cancer Genome Atlas data portal is an open access platform, and all data sets are available for download at https://tcga-data.nci.nih.gov/tcga/.

All studies received approval from their respective human research ethics committees, and details of TCGA can be found at http://cancergenome.nih.gov/.

## Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was partly supported by the Ministry of Science and Technology, R.O.C. (107-2221-E-214 -013 – and 106-2221-E-992 -327 -MY2), Taiwan.

## ORCID iD

Cheng-Hong Yang https://orcid.org/0000-0002-2741-0072

## Supplemental material

**Supplemental information** accompanies this paper at https://drive.google.com/open?id=1U50yuVHnX8moJclMx5qv-kZzJXmva-_o.

## References

1. Macher-Goeppinger S, Roth W, Wagener N, *et al.* Molecular heterogeneity of TFE3 activation in renal cell carcinomas. *Mod Pathol* 2012; 25: 308–315.
2. Zhao H, Ljungberg B, Grankvist K, *et al.* Gene expression profiling predicts survival in conventional renal cell carcinoma. *PLoS Med* 2006; 3: e13.
3. Mickisch GH. Principles of nephrectomy for malignant disease. *BJU Int* 2002; 89: 488–495.
4. Janzen NK, Kim HL, Figlin RA, *et al.* Surveillance after radical or partial nephrectomy for localized renal cell carcinoma and management of recurrent disease. *Urol Clin North Am* 2003; 30: 843–852.
5. Hakimi AA, Ostrovnaya I, Reva B, *et al.* Adverse outcomes in clear cell renal cell carcinoma with mutations of 3p21 epigenetic regulators BAP1 and SETD2: a report by MSKCC and the KIRC TCGA research network. *Clin Cancer Res* 2013; 19: 3259–3267.
6. Lee C-H and Motzer RJ. Immune checkpoint therapy in renal cell carcinoma. *Cancer J* 2016; 22: 92–95.
7. Yang C-H, Lin Y-D, Chuang L-Y, *et al.* MDR-ER: balancing functions for adjusting the ratio in risk classes and classification errors for imbalanced cases and controls using multifactor-dimensionality reduction. *PLoS One* 2013; 8: e79387.
8. Chen J-B, Chuang L-Y, Lin Y-D, *et al.* Preventive SNP–SNP interactions in the mitochondrial displacement loop (D-loop) from chronic dialysis patients. *Mitochondrion* 2013; 13: 698–704.
9. Chen J-B, Chuang L-Y, Lin Y-D, *et al.* Genetic algorithm-generated SNP barcodes of the mitochondrial D-loop for chronic dialysis susceptibility. *Mitochondrial DNA* 2014; 25: 231–237.
10. Chen J-B, Lee W-C, Cheng B-C, *et al.* Impact of risk factors on functional status in maintenance hemodialysis patients. *Eur J Med Res* 2017; 22: 54.
11. Yang C-H, Lin Y-D, Chuang L-Y, *et al.* Joint analysis of SNP–SNP-environment interactions for chronic dialysis by an improved branch and bound algorithm. *J Comput Biol* 2017; 24: 1212–1225.
12. Ricketts CJ, De Cubas AA, Fan H, *et al.* The cancer genome atlas comprehensive molecular characterization of renal cell carcinoma. *Cell Rep* 2018; 23: 3698.
13. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 2013; 499: 43–49.
14. Leung MK, Delong A, Alipanahi B, *et al.* Machine learning in genomic medicine: a review of computational problems and data sets. *Proc IEEE* 2015; 104: 176–197.
15. Hastie T, Tibshirani R and Friedman J. *The elements of statistical learning: data mining, inference, and prediction*. New York, NY: Springer Science & Business Media, 2009.
16. Murphy KP. *Machine learning: a probabilistic perspective*. Cambridge, MA: MIT Press, 2012.
17. Telenti A, Lippert C, Chang P-C, *et al.* Deep learning of genomic variation and regulatory network data. *Hum Mol Genet* 2018; 27: R63–R71.
18. Libbrecht MW and Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet* 2015; 16: 321–332.
19. Angermueller C, Pärnamaa T, Parts L, *et al.* Deep learning for computational biology. *Mol Syst Biol* 2016; 12: 878.
20. Zou J, Huss M, Abid A, *et al.* A primer on deep learning in genomics. *Nat Genet* 2019; 51: 12–18.
21. Nicolas E, Demidova EV, Iqbal W, *et al.* Interaction of germline variants in a family

with a history of early-onset clear cell renal cell carcinoma. *Mol Genet Genomic Med* 2019; 7: e556.

22. Beer DG, Kardia SL, Huang C-C*, et al*. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 2002; 8: 816–824.
23. Antonov A, Krestyaninova M, Knight R*, et al*. PPISURV: a novel bioinformatics tool for uncovering the hidden role of specific genes in cancer survival outcome. *Oncogene* 2014; 33: 1621.
24. Chen D-GD, Sun J and Peace KE. *Interval-censored time-to-event data: methods and applications*. Boca Raton, FL: Chapman and Hall/CRC, 2012.
25. Yang H, Zhang J, Yu B*, et al. Statistical methods for immunogenicity assessment*. Boca Raton, FL: Chapman and Hall/CRC, 2015.
26. Alarcón-Soto Y, Espasandín-Domínguez J, Guler I*, et al* . Data science in biomedicine. *arXiv Preprint arXiv*:190904486, 2019.
27. Linden A and Yarnold PR. Modeling time-to-event (survival) data using classification tree analysis. *J Eval Clin Pract* 2017; 23: 1299–1308.
28. Cox DR. *Analysis of survival data*. Boca Raton, FL: Chapman and Hall/CRC, 2018.
29. Wang P, Li Y and Reddy CK. Machine learning for survival analysis: a survey. *ACM Comput Surv* 2019; 51: 110.
30. Cox DR. Regression models and life-tables. *J R Stat Soc Ser B (Methodol)* 1972; 34: 187–202.
31. Katzman JL, Shaham U, Cloninger A*, et al*. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol* 2018; 18: 24.
32. Kvamme H, Borgan Ø and Scheel I. Time-to-event prediction with neural networks and Cox regression. *J Mach Learn Res* 2019; 20: 1–30.
33. Faraggi D and Simon R. A neural network model for survival data. *Stat Med* 1995; 14: 73–82.
34. Mariani L, Coradini D, Biganzoli E*, et al*. Prognostic factors for metachronous contralateral breast cancer: a comparison of the linear Cox regression model and its artificial neural network extension. *Breast Cancer Res Treat* 1997; 44: 167–178.
35. Xiang A, Lapuerta P, Ryutov A*, et al*. Comparison of the performance of neural network methods and Cox regression for censored survival data. *Comput Stat Data Anal* 2000; 34: 243–257.
36. Sargent DJ. Comparison of artificial neural networks with other statistical approaches: results from medical data sets. *Cancer* 2001; 91(Suppl. 8): 1636–1642.
37. LeCun Y, Bengio Y and Hinton G. Deep learning. *Nature* 2015; 521: 436–444.
38. Santurkar S, Tsipras D, Ilyas A*, et al*. How does batch normalization help optimization? *Adv Neural Inf Process Syst* 2018; 2483–2493.
39. Zupan B, DemšAr J, Kattan MW*, et al*. Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artif Intell Med* 2000; 20: 59–75.
40. Yousefi S, Amrollahi F, Amgad M*, et al*. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Sci Rep* 2017; 7: 11707.
41. Lee C, Zame WR, Yoon J*, et al*. Deephit: a deep learning approach to survival analysis with competing risks. In: *Proceedings of the AAAI conference on artificial intelligence*, New Orleans, LA, 2–7 February 2018, vol. 32, no. 1. Menlo Park: AAAI.
42. Srivastava N, Hinton G, Krizhevsky A*, et al*. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014; 15: 1929–1958.
43. Finley LW, Zhang J, Ye J*, et al*. SnapShot: cancer metabolism pathways. *Cell Metab* 2013; 17: 466.e2.
44. Jones RG and Thompson CB. Tumor suppressors and cell metabolism: a recipe for cancer growth. *Genes Dev* 2009; 23: 537–548.
45. ter Braak B, Wink S, Koedoot E*, et al*. Alternative signaling network activation through different insulin receptor family members caused by pro-mitogenic antidiabetic insulin analogues in human mammary epithelial cells. *Breast Cancer Res* 2015; 17: 97.
46. Davis AJ and Chen DJ. DNA double strand break repair via non-homologous end-joining. *Transl Cancer Res* 2013; 2: 130–143.
47. Maeda Y, Farina NH, Matzelle MM*, et al*. Synovium-derived microRNAs regulate bone pathways in rheumatoid arthritis. *J Bone Miner Res* 2017; 32: 461–472.
48. Hoshino Y, Sonoda H, Nishimura R*, et al*. Involvement of the NADPH oxidase 2 pathway in

renal oxidative stress in Aqp11-/- mice. *Biochem Biophys Rep* 2019; 17: 169–176.

49. Baheti S, Singh P, Zhang Y*, et al*. Adipose tissue DNA methylome changes in development of new-onset diabetes after kidney transplantation. *Epigenomics* 2017; 9: 1423–1435.

50. Javadneia A, Moghadam MT, Alivand A*, et al*. Sperm DNA damage before and after kidney transplantation. *Nephro-Urol Mon* 2019; 11: e86990.

51. De La Rocha AMA, Lopez-Bertoni H, Guruceaga E*, et al*. Analysis of SOX2-regulated transcriptome in glioma stem cells. *PLoS One* 2016; 11: e0163155.

52. Trivedi S and Starz-Gaiano M. Drosophila Jak/STAT signaling: regulation and relevance in human cancer and metastasis. *Int J Mol Sci* 2018; 19: 4056.

53. Chuang PY and He JC. JAK/STAT signaling in renal diseases. *Kidney Int* 2010; 78: 231–234.

54. Heng DY, Xie W, Regan MM*, et al*. External validation and comparison with other models of the international metastatic renal-cell carcinoma database consortium prognostic model: a population-based study. *Lancet Oncol* 2013; 14: 141–148.

55. Paszke A, Gross S, Chintala S*, et al Automatic differentiation in pytorch*. Long Beach, CA: NIPS, 2017.

56. Ioffe S and Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv Preprint arXiv*:150203167, 2015.

57. Kingma DP and Ba J. Adam: a method for stochastic optimization. *arXiv Preprint arXiv*:14126980, 2014.

58. Kaplan EL and Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958; 53: 457–481.

59. Harrell FE Jr, Lee KL, Califf RM*, et al*. Regression modelling strategies for improved prognostic prediction. *Stat Med* 1984; 3: 143–152.

60. Ashburner M, Ball CA, Blake JA*, et al*. Gene ontology: tool for the unification of biology. *Nat Genet* 2000; 25: 25.

61. Subramanian A, Tamayo P, Mootha VK*, et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005; 102: 15545–15550.