Supplementary

Appendix 1 Whole exome sequencing

The original fluorescence image files derived from the Illumina platform undergo transformation into short reads (raw data) through base calling, which is then recorded as a FASTQ format. The process of quality control comprised the following steps: (I) Getting rid of paired reads that show adapter contamination, which means that more than 10 nucleotides align to the adapter, allowing for ≤10% mismatches; (II) Eliminating paired reads with a significantly high rate, over 10%, of uncertain bases; (III) discharging paired reads with a low quality (Phred quality <5) base, surpassing 50%. After ensuring a clean, high-quality sequencing data, it's mapped to the reference genome (GRCh38) using the Burrows-Wheeler Aligner (BWA) software (40) to procure the original mapping result in the BAM format. In the subsequent steps, software tools like SamTools (41) and Sambamba are utilized for sorting bam files and marking duplicates to create the final bam file. Variant calling and identification of SNP and InDels are carried out using Samtools (41) mpileup and bcftools. The detection of somatic SNV is achieved through muTect, somatic InDel by Strelka, and somatic CNV is determined using Control-FREEC (42). The annotation is performed using ANNOVAR (43) for the VCF (Variant Call Format) file obtained from the previous steps. Detailed information about variant position, variant type, and conservative prediction is retrieved using multiple databases, including dbSNP, 1,000 Genome, esp6500, GnomAD, CADD, HGMD, and COSMIC, etc. In the quest to identify exonic variants, gene transcript annotation databases like Consensus CDS, RefSeq, Ensemble and UCSC are implemented to determine amino acid alterations. Functional annotation is performed using Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, and Biocarta.

References

40. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754-60.
41. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078-9.
42. Boeva V, Popova T, Bleakley K, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics 2012;28:423-5.
43. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:e164.

https://dx.doi.org/10.21037/tlcr-23-737

Table S1 Distribution of train and validation patient samples of multi-centers

	Training cohort	Validation cohort	Total
Hospital 1	97	8	105
Hospital 2	16	8	24
Hospital 3	16	2	18
Hospital 4	10	11	21
Hospital 5	22	1	23
Hospital 6	11	11	22
Hospital 7	\	6	6
Hospital 8	5	5	10
Hospital 9	\	24	24
Hospital 10	\	2	2
Hospital 12	\	11	11
Hospital 13	\	5	5
Hospital 14	\	2	2
Hospital 15	\	2	2
Hospital 16	\	3	3
Hospital 17	\	5	5
Total	177	106	283

Table S2 Distribution of external validation patient samples of multi-centers

Center code	External validation cohort
101	13
102	3
106	4
107	5
108	1
113	5
114	4
116	1
117	3
121	6
124	1
125	1
Total	47

https://dx.doi.org/10.21037/tlcr-23-737

Table S3 List of features after data preprocessing

Feature	Description	Feature	Description
Age	Age of patient	Ccr	Creatinine clearance
BMI	Body mass index	D-dimer	Concentration of D-Dimer
ALT	Concentration of alanine aminotransferase	CK	Creatine kinase activity
AST	Concentration of Aspartate aminotransferase	α-HBDH	Concentration of α-hydroxybutyrate dehydrogenase
ALT/AST	the ratio between the concentrations of the enzymes AST and ALT	PT	Prothrombin time
ALP	Concentration of alkaline phosphatase	APTT	Activated partial thromboplastin time
LDH	Concentration of lactate dehydrogenase	Hbc	Concentration of haemoglobin
TP	Concentration of total protein	RBC	Concentration of red blood cells
ALB	Concentration of albumin	WBC	Concentration of white blood cells
TBIL	Bilirubin	NeuA	Concentration of neutrophils
UrBunIL	Concentration of urea nitrogen	LymA	Concentration of lymphocyte
Crea	Concentration of creatinine	MoA	Concentration of monocyte
Glu	Concentration of glucose	Platelet	Count of platelets
K	Concentration of potassiun	uPH	Urine pH
NA	Concentration of sodium	Css_{max}	peak concentration at steady-state
CL	Concentration of chloride	Css_{min}	steady-state plasma concentrations at the trough
Ca	Concentration of calcium	Sex = male	0 for female; 1 for male
Mg	Concentration of magnesium	Smoking = yes	0 for not smoking; 1 for smoking.

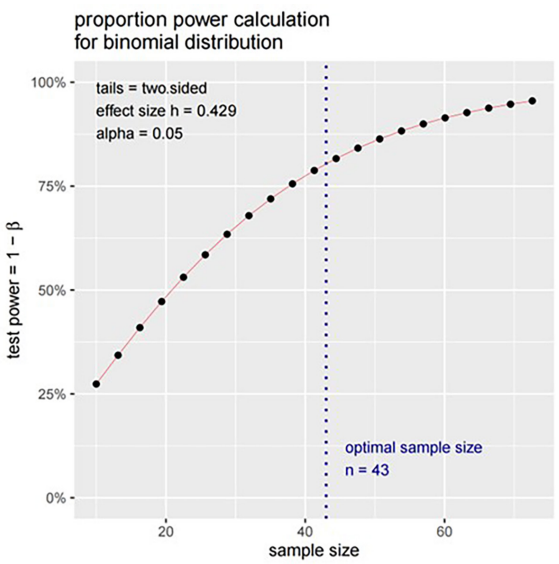

Figure S1 The plot of sample size and power.

https://dx.doi.org/10.21037/tlcr-23-737

Table S4 Correlation of the most correlated features

Rank	Feature 1	Feature 2	Pearson correlation coefficient
1	NeuA	WBC	0.9539
2	Css_{min}	Css_{max}	0.8377
3	LDH	α-HBDH	0.8207
4	Hbc	RBC	0.7110
5	WBC	MoA	0.6841
6	Sex = male	Smoking = yes	0.6786
7	NeuA	MoA	0.6169
8	NA	CL	0.6158
9	ALT	AST	0.5933
10	Crea	Sex = male	0.5334
11	PT	APTT	0.5055

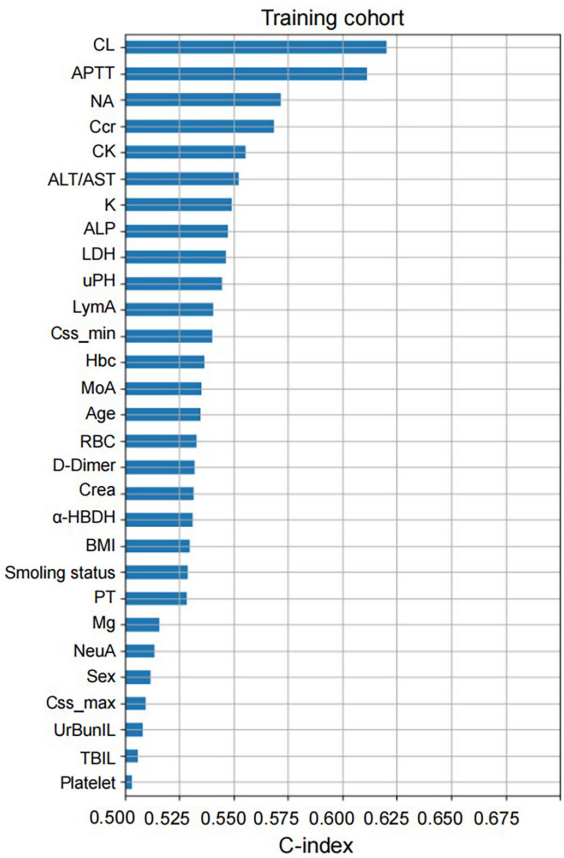

Figure S2 Feature importance analysis and selection. The C-index of each feature model individually in training cohort by CoxNet.

https://dx.doi.org/10.21037/tlcr-23-737

Table S5 Performance comparison of different models in survival analysis using the selected 4 features

Method	Cross-validation (averaged)
CoxNet	0.6443
CoxSVM	0.6663
DeepSurv	0.6681
CoxMoE	0.6761

Table S6 Performance comparison of different deep-learning models in multi-task modeling

Method	DeepSurv	CoxMoE
Cross-validation (averaged)		
Risk score (prediction, C-index)	0.6527	0.6732
Treatment response (prediction, ACC)	0.7564	0.7714
Treatment response (prediction, AUC)	0.7814	0.8181

ACC, accuracy; AUC, area under the curve.

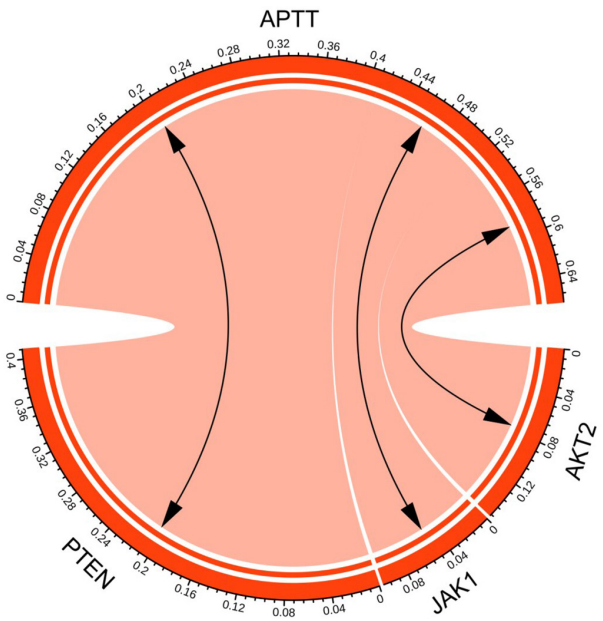

Figure S3 The correlation between APTT and three representative genes that were involved in EGFR-TKI drug resistance. APTT, activated partial thromboplastin time.

https://dx.doi.org/10.21037/tlcr-23-737