TYPE Methods
PUBLISHED 15 August 2024
DOI 10.3389/fgene.2024.1381851

SAFE-MIL: a statistically interpretable framework for screening potential targeted therapy patients based on risk estimation

Yanfang Guan[1,2,3†], Zhengfa Xue[1,2†], Jiayin Wang[1,2†], Xinghao Ai[4†], Rongrong Chen[3], Xin Yi[3], Shun Lu[4]* and Yuqian Liu[1,2]*

[1]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China, [2]Shaanxi Engineering Research Center of Medical and Health Big Data, Xi'an Jiaotong University, Xi'an, China, [3]Geneplus Beijing Institute, Beijing, China, [4]Shanghai Chest Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Check for updates

OPEN ACCESS

EDITED BY
Xiao Li,
Shandong Provincial Qianfoshan Hospital, China

REVIEWED BY
Zhongheng Zhang,
Sir Run Run Shaw Hospital, China
Zhangyang Ai,
Hunan University, China

*CORRESPONDENCE
Yuqian Liu,
yuqian.liu@xjtu.edu.cn
Shun Lu,
shunlu@sjtu.edu.cn

†These authors have contributed equally to this work and share first authorship

RECEIVED 04 February 2024
ACCEPTED 31 July 2024
PUBLISHED 15 August 2024

CITATION
Guan Y, Xue Z, Wang J, Ai X, Chen R, Yi X, Lu S and Liu Y (2024) SAFE-MIL: a statistically interpretable framework for screening potential targeted therapy patients based on risk estimation.
Front. Genet. 15:1381851.
doi: 10.3389/fgene.2024.1381851

Patients with the target gene mutation frequently derive significant clinical benefits from target therapy. However, differences in the abundance level of mutations among patients resulted in varying survival benefits, even among patients with the same target gene mutations. Currently, there is a lack of rational and interpretable models to assess the risk of treatment failure. In this study, we investigated the underlying coupled factors contributing to variations in medication sensitivity and established a statistically interpretable framework, named SAFE-MIL, for risk estimation. We first constructed an effectiveness label for each patient from the perspective of exploring the optimal grouping of patients' positive judgment values and sampled patients into 600 and 1,000 groups, respectively, based on multi-instance learning (MIL). A novel and interpretable loss function was further designed based on the Hosmer-Lemeshow test for this framework. By integrating multi-instance learning with the Hosmer-Lemeshow test, SAFE-MIL is capable of accurately estimating the risk of drug treatment failure across diverse patient cohorts and providing the optimal threshold for assessing the risk stratification simultaneously. We conducted a comprehensive case study involving 457 non-small cell lung cancer patients with EGFR mutations treated with EGFR tyrosine kinase inhibitors. Results demonstrate that SAFE-MIL outperforms traditional regression methods with higher accuracy and can accurately assess patients' risk stratification. This underscores its ability to accurately capture inter-patient variability in risk while providing statistical interpretability. SAFE-MIL is able to effectively guide clinical decision-making regarding the use of drugs in targeted therapy and provides an interpretable computational framework for other patient stratification problems. The SAFE-MIL framework has proven its effectiveness in capturing inter-patient variability in risk and providing statistical interpretability. It outperforms traditional regression methods and can effectively guide clinical decision-making in the use of drugs for targeted therapy. SAFE-MIL offers a valuable interpretable computational framework that can be applied to other patient stratification problems, enhancing the precision of risk assessment in personalized medicine. The source code for SAFE-MIL is available for further exploration and application at https://github.com/Nevermore233/SAFE-MIL.

KEYWORDS

EGFR, non-small cell lung cancer, target therapy, risk estimation, Hosmer-Lemeshow test, multi-instance learning

1 Introduction

Lung cancer holds the highest incidence rate among all types of cancer and boasts the largest selection of approved targeted therapeutic agents. Consequently, targeted therapy for lung cancer has become a prevalent and routine treatment in clinical practice. However, patients with target mutations show heterogeneous and diminished responses to the treatment (Cheng et al., 2020). Several studies have shown that its effectiveness is related to the gene mutation abundance level (Blakely et al., 2017; Samstein et al., 2019). Higher mutation abundance level may be related to better treatment outcomes (Zhou et al., 2011; Yan et al., 2019; Wang et al., 2021; Liu et al., 2022), such as a longer median survival time (Liu et al., 2022). In the event of a low abundance level, the therapeutic effect of targeted drugs may be reduced or completely lost (Tang et al., 2021). However, this issue remains unresolved in clinical practice. It is due to the absence of statistically interpretable tools that can accurately assess the risk of treatment failure in patients receiving targeted therapy. Furthermore, the optimal stratification threshold for targeted therapy based on mutation abundance level is unknown among the patient cohort. Hence, there is an urgent and unmet clinical need to develop a broadly applicable and statistically interpretable method approach capable of effectively assessing the risk associated with medicine utilization and then give the optimal threshold to assist clinical decision-making in target therapy.

Performing drug screening and selecting appropriate personalized treatment based on individual genomic, proteomic, and clinical features is one of the paramount goals of precision medicine (Nemati et al., 2018; Zhang et al., 2020; Banerjee et al., 2021; Sotudian and Paschalidis, 2022; Diao et al., 2023; Chen et al., 2024). Various clinical trials generate high quality results on effectiveness comparison of different drugs for the treatment of the same disease which is crucial for drug development and clinical practice (Rubin and Gilliland, 2012; Zhang Y. et al., 2021). Clinically, patients harboring same target gene mutations have different medication risks due to differences in mutation abundance level and drug sensitivity (Robichaux et al., 2021; Wang et al., 2021). The use of inappropriate drug regimens in patients with high risk may lead to delays in the patient's condition and additional medical costs. One of the results is that mild patients may suffer severe or even catastrophic lesions (Schnipper et al., 2015; Daoud et al., 2020). Therefore, it is very necessary to assess the risk of medication use and utilize it in the clinical.

Nevertheless, this problem is different from the conventional drug effectiveness prediction problem. With the rapid development of artificial intelligence, a multitude of computational approaches to drug effectiveness prediction have been developed (Göttlich et al., 2016; Chang et al., 2022; Peng et al., 2022; Łosińska et al., 2022; Yang et al., 2023). Most of these approaches focus on utilizing the powerful feature extraction abilities and learning capabilities of artificial intelligence models to predict the drug responses of patients. Various computational models were employed to obtain potential vector representations of drugs and diseases for drug effectiveness prediction. However, a deep learning model is a black box, posing challenges in elucidating the acquired knowledge and the underlying principles guiding its predictive capabilities (Kuenzi et al., 2020). The existing drug effectiveness prediction models exhibit a tendency towards excessive complexity and lack relevant statistical interpretations, such as deep neural networks. According to Food and Drug Administration (FDA) policy, the assessment needs to be conducted in a manner that allows for statistical interpretation. Hence, this is a risk likelihood estimation problem rather than a prediction problem. Furthermore, the patient stratification problem is a matter of interest for both the clinical and drug regulatory communities, rather than the drug effectiveness prediction. Therefore, these existing drug effectiveness prediction models have high predictive accuracy; yet, their practical applicability in clinic practice is limited. Exploring the challenge of designing a reasonable model from the perspective of optimal patient stratification to consider the risk of treatment failure holds significant academic value.

The widespread application of multiple instance learning (MIL) in drug-target, drug activity, and drug effectiveness prediction (Fu et al., 2012; Zhao et al., 2013; Saberian et al., 2022; Wang et al., 2022) has attracted our attention. In clinical practice, the risk of drug failure is typically assessed on a population-wide scale, making it challenging to provide a personalized estimate of treatment failure risk for an individual patient prior to the initiation of medication. In MIL, the dataset is divided into groups of multiple instances, with the group labels known but the instance labels unknown, which has a good correlation with patient grouping in clinical drug practice. For example, suppose there are many microscopic tissue slice images used to detect cancer. Each image may contain multiple small regions, which we refer to as "instances." The whole image can be labeled as "cancer" or "no cancer," but it may be uncertain whether each specific small region contains cancer cells. In this case, we can consider each image as a "group" containing multiple "instance regions." The algorithm needs to predict the label of the "group" based on the known labels of the "instances." In the context of drug effectiveness prediction, using MIL to consider and model drug effectiveness may be a feasible idea.

This paper presents a generic novel framework termed SAFE-MIL for screening potential targeted therapy patients based on risk estimation, using epidermal growth factor receptor (EGFR)-mutant non-small cell lung cancer (NSCLC) as an example. The proposed framework integrates multi-instance learning with the Hosmer-Lemeshow test (Hosmer and Lemesbow, 1980) to estimate the

TABLE 1 Demographic and baseline characteristics of cohort-1,2,3.

Participants (n = 457)						
	Cohort1(n = 100)		Cohort2 (n = 237)		Cohort3(n = 120)	
Gender	NO.	%	NO.	%	NO.	%
Female	64	64	142	60	68	57
Male	36	36	95	40	52	43
Age (26–86)						
>60	56	56	111	47	58	48
<=60	44	44	126	53	62	52
Stage						
III	10	10	0	0	0	0
IV	90	90	237	100	120	100

risk of drug use, which provides statistical interpretability and can address the limitations of traditional methods. Experimental results demonstrate that SAFE-MIL provides interpretability and higher accuracy, hence potentially facilitating the process of therapeutic drug selection for patients.

2 Patients and methods

In this paper, we present a spontaneous formulation of the drug risk estimation problem as an instance of multi-instance learning. We address the problem in three stages, first organizing the given clinical information and drug action information of patients into instances, then identifying the estimated scores of all instances in the same group, and finally combining all the estimated scores as the output. Prior to these stages, data preprocessing was performed.

2.1 Patients and samples collection

In this research, we acquired three sets of Non-Small Cell Lung Cancer (NSCLC) sequencing data, exclusively sourced from EGFR-mutated NSCLC patients. Inclusion criteria for patients in this study: Patients must carry EGFR mutations (EGFRm) and be clinically selected for anti-tumor treatment with a tyrosine kinase inhibitor (TKI) based on the detection results. Exclusion criteria is that patients who have not received EGFR-TKI treatment according to the detection results, or patients carrying mutations in other drug target sites. The first batch consists of data from 100 patients with treatment-naive NSCLC (stage III–IV), all of whom underwent first-line targeted therapy. This information was gathered from 14 medical centers spanning the period from 23 February 2017, to 31 December 2019. The trial has been officially registered with the identifier NCT03059641. The second and third batches comprises data from 237 patients and 120 patients (stage IV) with EGFR-mutated NSCLC who underwent next-generation sequencing at Geneplus-Beijing (Beijing, China) between 2016 and December 2019 (Table 1). Tumor tissue samples were utilized to identify actionable mutations for targeted therapy in these patients. Notably, all these patients received anti-EGFR targeted therapy. It’s important to note that all participants in the study provided written informed consent. The research was conducted in accordance with a protocol approved by the Institutional Review Board of Shanghai Chest Hospital. Cohort 1 is the first batch of NSCLC patients were collected. In order to better generate multi-instance datasets, we conducted experiments using cohort 1 as independent data.

Cohort 3 comprised 120 patients with NSCLC accepted EGFR-TKI target therapy (43.3% male, stage IV) (Supplementary Table S5), with ages at diagnosis ranging from 35 to 83 years. In the tumor samples, a total of 952 SVs are included (Supplementary Table S6). All patients were found to carry EGFR-sensitive mutations. Cohorts 1 and 2, 3 have recorded slightly different survival information, with the former documenting patients’ progression-free survival (PFS), while the latter records time to treatment failure (TTF).

DNA extraction, targeted capture, and NGS Genetic analysis for all data were performed as previously described (Nong et al., 2018; Zhang et al., 2019). Sequencing libraries were prepared from genomic DNA were prepared using Illumina TruSeq DNA Library Preparation Kits (Illumina) or MGIEasy Universal Library Prep Set (MGI Tech). Libraries were hybridized to custom-designed biotinylated oligonucleotide probes (Integrated DNA Technologies, Inc) targeting 1,021 genes. Prepared libraries were sequenced on a NextSeq CN 500 (Illumina) or MGISEQ-2000 sequencer (MGI Tech, Shenzhen, China). After the entire run was completed, image analyses, error estimation and base calling were performed to generate primary data. We then removed a few unqualified sequences from the primary data using a local dynamic programming algorithm, which included low-quality reads, defined as reads that contained more than 10 percent Ns in the read length, 50% reads with a quality value of less than five and with an average quality of less than 10 and adapter sequences including indexed sequence. The remaining sequences were termed as clean reads for further analysis. The clean reads were aligned to the reference human genome (hg19) with Burrows-Wheeler Aligner (version 0.7.12-r1039). Variants were called with GATK (version 3.4–46-gbc02625) and MuTect (version 1.1.4). Contra (v2.0.8) was used to detect copy-number variants (CNVs), and NCsv (in-house software version 0.2.3) was used to detect structural variants (SVs). Targeted capture sequencing required a minimal mean effective depth of coverage of 300 in tumor tissues Variants were filtered to exclude synonymous variants, known germline variants in dbSNP, and variants that occur at a population abundance level of >1% in the Exome Sequencing Project.

2.2 The model and algorithm

The SAFE-MIL framework was used, which utilizes multiple instances learning to predict drug failure risk in patients. The main steps of the SAFE-MIL framework include:

i. Designing drug effectiveness labels based on clinical features of patients.
ii. Building a model using multiple instances learning to predict drug failure risk by aggregating the labels of each patient.
iii. Using a new loss function based on the Hosmer-lemeshow test to capture differences between groups and improve the performance and statistical interpretability of the model.
iv. Learn the relationship between mutation abundance level and drug failure risk, and determine the optimal positive threshold for drug failure.

Cohort 1 is the first batch of NSCLC patients we collected. In order to better generate multi-instance datasets, we conducted experiments using cohort 1 as independent data. We randomly divided cohort 1 into a training set and a testing set and conducted independent experiments. Similarly, we subsequently collected more samples (cohort 2 and cohort 3) and conducted similar independent experiments. Due to slight differences in collection time, clinical features, and other aspects among these three cohorts, we did not merge them.

2.3 The design of drug effectiveness labels for patients

In order to better illustrate the variations in therapeutic efficacy among different patients, we have considered representing the therapeutic efficacy of a drug as a probability value ranging from 0 to 1. Initially, we evaluated the correlation between patients' clinical characteristics and the effectiveness of the drug based on existing literature. Given that a patient's drug response is closely linked to their PFS (Paz-Ares et al., 2018), we utilized the patients' PFS to determine the drug's effectiveness for each individual of cohort 1. For cohort 2, we attempted to assess the efficacy of the drug using TTF. To begin, patients with a PFS/TTF exceeding 8 months as of the cut-off value were included in the experimental cohort. Subsequently, we employed a min-max scaling technique to map the PFS/TTF values of these patients onto a scale of 0–1. This scaled value was then used to represent the probability of the patient's drug efficacy.

2.4 Data preprocessing and patient grouping

The data preprocessing process is shown in Figure 2A. Given that the efficacy of a medication for an individual patient remains unobservable prior to treatment, we propose an unsupervised clustering approach to categorize patients with analogous characteristics into distinct cohorts. Subsequently, we employ multi-instance learning to estimate the probability of drug effectiveness across these patient groups. We used the K-means unsupervised clustering algorithm to group patients into four clusters (details see Supplementary Materials S1, S2, Figure S1). We then randomly sampled 1–10 patients with replacement to form a group, with each patient serving as an instance. We created datasets based on the two aforementioned patient cohorts. For each cohort, we generated two datasets of varying sizes to demonstrate the model's generalization capacity. One dataset comprised 600 bags (cohort 1–600 and cohort 2–600), with 150 bags randomly generated from each cluster. The other dataset included 1,000 bags (cohort 2–1,000 and cohort 2–1,000), with 250 bags randomly generated from each cluster. The drug effectiveness of the groups was determined by calculating the average value of the drug effectiveness label assigned to the instances in the group. The raw input data, preprocessed output data, and grouped data can be accessed at GitHub repository.

2.5 Objective function

Here, risk estimation is formulated as a regression problem. We combine multiple instances learning and Hosmer-lemeshow test to design our model. Suppose that the training set consists of N training groups $\{G_1, G_2, \ldots, G_N\}$ and each group contains M_i instances $\{G_{i1}, G_{i2}, \ldots, G_{iMi}\}$, where the label corresponding to each training group G_i is marked L_i. In addition, each instance in the groups is a p-dimensional attribute value vector.

In multi-instance learning, the task is to predict the labels of unseen groups. The learning algorithm can acquire the label of the group, but cannot obtain the label of the instance. Therefore, we define the global error function of the neural network at the group level using the labels of the training groups as

$$E = \sum_{i=1}^{N} E_i \tag{1}$$

Where E_i is the output error corresponding to the group G_i (Equation 1).

To enhance the statistical interpretability of our framework and establish trust between clinicians and our framework, we consider designing loss functions based on statistical tests. We found that the Hosmer-Lemeshow test (Hosmer and Lemesbow, 1980), which evaluates the goodness-of-fit of grouped data, is closely related to multiple instance learning. The Hosmer-Lemeshow test is widely used in the evaluation of risk models (Kramer and Zimmerman, 2007; Zhang L. et al., 2021; Davies et al., 2022). Given the excellent performance of the Hosmer-Lemeshow test in model verification, we design the loss function according to the calculation formula of Hosmer-Lemeshow (HL) statistic (HL_s) (Equation 2). The Hosmer-Lemeshow test is a statistical method used to evaluate the goodness-of-fit for binary logistic regression models. The basic idea is to divide the data into several groups and then compare the actual observed values with the predicted values in each group. In the Hosmer-Lemeshow test, the p-value is calculated based on the chi-squared distribution, corresponding to the Hosmer-Lemeshow statistic's cumulative distribution function value. If the p-value is very small (less than 0.05), it indicates that the model does not fit well, meaning there is a significant difference between the predicted probabilities and the actual observed values. If the p-value is large, it indicates that the model fits well, meaning there is no significant difference between the predicted probabilities and the actual observed values. Suppose that the observed value (*Observed.A*, *Observed.not.A*) and expected values

(*Expected.A*, *Expected.not.A*) of *Event A* are divided into Q groups $\{G_1, G_2, \ldots, G_Q\}$, the HL statistic and *p*-value (HL_p) of HL test are defined as

$$HL_s = \sum_{q=1}^{Q}\left(\frac{(Observed.A - Expected.A)^2}{Expected.A} + \frac{(Observed.not.A - Expected.not.A)^2}{Expected.not.A}\right) \tag{2}$$

$$HL_p = 1 - chisq(HL_s, Q-2) \tag{3}$$

where *chisq* (HL_s, *Q*-2) is the chi-square distribution with *Q*-2 degrees of freedom (Equation 3 more details in Supplementary Material S3).

Since the HL based loss function (HL loss) is partially non-differentiable, we use the mean square error loss function to replace the non-differentiable points of the HL loss function (Equation 4). Research indicates that the output value of a group in neural network is determined by the maximum output value of the instance in the group (Amar et al., 2001). Therefore, to simulate the above rules, the output error of group G_i is defined as

$$E_i = \begin{cases} \dfrac{\left(max(o_{ij}) - L_i\right)^2}{\max\limits_{0\le j\le M_i}(o_{ij})} & \max\limits_{0\le j\le M_i}(o_{ij}) \neq 0 \\ \dfrac{1}{2}\left(\max\limits_{1\le j\le M_i}(o_{ij}) - L_i\right)^2 & \max\limits_{0\le j\le M_i}(o_{ij}) = 0 \end{cases} \tag{4}$$

where o_{ij} is the network output corresponding to the instance G_{ij}.

The input of the model is all the training groups, and the output is the drug effective probability of the patients in the group, which is used to measure the patient's drug use risk (1- drug effective probability). The hyperparameters used for SAFE-MIL are summarized in Table 2. In addition, we explored the correlation between the risk stratification and the survival benefits of the patients. In all analyses, $P < 0.05$ was considered statistically significant.

In addition, after training the model, we search for the optimal positive threshold that can differentiate the drug use risk of patients based on the model's outputs on the training set. Firstly, we define the search space for the optimal positive threshold as [MA_{margin}- *margin*, + MA_{margin} *margin*], where MA represents the mutation abundance and the margin boundaries determine the size of the search space (Equation 5). The objective function of optimal positive threshold is defined as follows:

$$diff = \max_t \left| \frac{1}{j-t}\sum_{i=t}^{j} d_i - \frac{1}{t}\sum_{i=1}^{t} d_i \right|$$
$$\tau = T_t \tag{5}$$

Where d_i represents the drug failure risk, and τ represents the optimal positive threshold. When t is set to T_t, the samples are divided into two groups based on τ, and the average drug failure risk difference between them is the largest.

TABLE 2 Hyperparameter of the SAFE-MIL.

Hyper-parameters	Setting
Learning rate	0.001
Epochs	200
Optimizer	Relu
Hidden layer neuron	50

2.6 Baselines

We evaluated SAFE-MIL with the four classical regression loss functions, including mean absolute error loss (MAE loss) (Fisher, 1915) (Equation 6), mean square error loss (MSE loss) (Equation 7) (Fisher, 1921), Huber loss (Equation 8) (Huber, 1992) and Log-cosh loss (Equation 9) (Shen et al., 2018). During one neural network iteration, the formulas of the four regression loss functions are defined as

$$Mse\,loss = \frac{1}{2}\left(y_p - y_t\right)^2 \tag{6}$$

$$Mae\,loss = \left|y_p - y_t\right| \tag{7}$$

$$Huber\,loss = \begin{cases} \frac{1}{2}\left(y_p - y_t\right)^2 & for\left|y_p - y_t\right| \le \delta \\ \delta\left|y_p - y_t\right| - \frac{1}{2}\delta^2 & otherwise \end{cases} \tag{8}$$

$$Log\cosh loss = \log\left(cosh\left(y_p - y_t\right)\right) \tag{9}$$

where y_p is the predicted value of the group, and y_t is the true value of the group. In Huber loss, the value of $\boldsymbol{\delta}$ is determined by cross-validation. In addition to evaluating the model using the mean square error (MSE), we also use the HL statistic and its corresponding *p*-value to test the accuracy and rationality of the predicted value. In the HL test, if the *p*-value is greater than 0.05, it means that the model has passed the HL test, which means that there is no significant difference between the predicted value and the real value, otherwise it means that the model fit is poor. The complete calculation process of the risk based on the SAFE-MIL framework is given in Algorithm 1.

```
Input: N training groups{$G_1$, $G_2$,...,$G_N$}.
Output: Drug failure risk of N groups {$D_1$, $D_2$, ... , $D_N$},
Q1, Q2, Q3.
SAFE-MIL (Epochs, Threshold)
    Initialize neural network Net;
    for (epoch=1; epoch<=Epochs; epoch++)
         GlobalErr=0; //Set the initial value of global
        error to be zero
         for (i=1; i<=N; i++)
            Compute the output error $E_i$ of group $B_i$
            according to Equation 2;
            GlobalErr = GlobalErr+$E_i$;
            The weights in Net are modified according to
            $E_i$ and the
            weight-updated rule of BP algorithm;
         end
         If (GlobalErr<=Threshold)
             return Net;
         end
    end
    return Net;
```

Algorithm 1. SAFE-MIL.

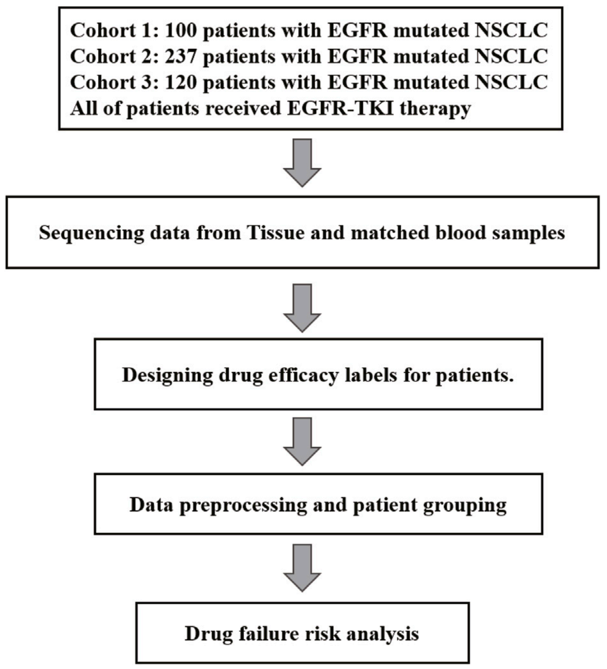
Cohort 1: 100 patients with EGFR mutated NSCLC
Cohort 2: 237 patients with EGFR mutated NSCLC
Cohort 3: 120 patients with EGFR mutated NSCLC
All of patients received EGFR-TKI therapy
Sequencing data from Tissue and matched blood samples
Designing drug efficacy labels for patients.
Data preprocessing and patient grouping
Drug failure risk analysis

FIGURE 1
Flow of participants in the study.

3 Results

3.1 Patient characteristics and somatic variation detection

Cohort 1 included 100 patients with NSCLC accepted first-line EGFR-TKI target therapy (36% male, 90% stage IV) (Supplementary Table S1). The age at diagnosis ranged from 33 to 80 years, with a median of 61 years (Table 1). In the tumor samples, a total of 592 somatic alterations were detected, including 539 SNVs (single nucleotide variants, SNVs) and small indels (insertion and deletions, indels), 50 copy number variants (CNVs), and 1 structural variants (SVs) (Supplementary Table S2). Cohort 2 comprised 237 patients with NSCLC accepted EGFR-TKI target therapy (40% male, stage IV) (Supplementary Table S3), with ages at diagnosis ranging from 26 to 86 years and a median age of 60 years (Table 1). In the tumor samples, a total of 1738 somatic alterations were detected, including 1,454 SNVs, 268 CNVs, and 4 SVs (Supplementary Table S4). Cohort 3 comprised 120 patients with NSCLC accepted EGFR-TKI target therapy (43.3% male, stage IV) (Supplementary Table S5), with ages at diagnosis ranging from 35 to 83 years (Table 1). In the tumor samples, a total of 930 SNVs or Indels, and 12 structural variants (SVs) were detected (Supplementary Table S6). All patients were found to carry EGFR-sensitive mutations. Cohorts 1 and 2, 3 have recorded slightly different survival information, with the cohort one documenting patients' progression-free survival (PFS), while the cohort 2/3 records time to treatment failure (TTF). In our study, both PFS and TTF were used as evaluation criteria to assess the predictive performance of our algorithm. These two outcomes are related, they showed a high correlation in NSCLC of targeted therapy, immunotherapy, and chemotherapy. Especially in targeted therapies of EGFRm-TKI, the correlation is even higher (r = 0.91, 95% CI 0.90,0.92) (Blumenthal et al., 2019).While PFS focuses specifically on disease progression, TTF offers a broader perspective by considering various factors influencing treatment continuation. The study utilized clinical data from patients and somatic alterations for the SAFE-MIL investigation. The study workflow is illustrated in Figure 1.

3.2 The result of SAFE-MIL for predicting EGFR-TKI failure risk in patients with EGFR-mutated NSCLC

The workflow of this study is shown in Figure 2B. We collected clinical data and SNV mutation data of 457 EGFR-mutated NSCLC patients and conducted data preprocessing. Following the mentioned criteria from the preceding text, we calculated the drug effectiveness label for each patient. Subsequently, an unsupervised clustering of patients was performed using the k-means algorithm based on the features including age, gender, mutation abundance, TP53 co-mutated or not, and the number of co-mutations, which were outlined in Table 3. Consequently, we normalized these features using min-max normalization and employed them for unsupervised clustering using the k-means algorithm. After the feature selection process, mutation abundance was identified as the most significant feature and selected as a key predictor. Although the presence of co-TP53 mutation and the number of co-mutations are two important genomic molecular features, but in this study the two features were indicated not important. The reasons may be from two aspects: firstly, the enrolled population was exclusively EGFR-mutant, with a relatively homogeneous molecular subtype; secondly, the data used in this study originated from a panel sequencing of ~ 1Mb, which may have limited the representation of the feature regarding the number of co-mutations compared to whole-exome sequencing data. Finally, we randomly selected 1 to 10 patients from each cluster using a with-replacement sampling method to generate groups. The drug effectiveness label of each group was determined by the average effectiveness label of the instances within the group.

We compared the performance of the HL loss function and four traditional regression loss functions in terms of mean squared error (MSE), HL statistic, and *P*-value. In the cohort1, overall, all loss functions passed the HL test with a *p*-value greater than 0.05 (Figure 3A), and achieved low mean squared error (MSE) (Figure 3B). SAFE-MIL demonstrated the lowest MSE and HL statistics among the two generated datasets in predicting drug effectiveness (Figures 3B, C), indicating that the HL-based loss functions can capture differences among instances within groups, ensuring accurate estimation results while maintaining statistical interpretability. Similarly, in the cohort2, SAFE-MIL based on the HL loss also achieved lower MSE (Figure 4). Due to the larger sample size of cohort2 compared to cohort1, theoretically, the model of HL-based SAFE-MIL predictive performance should be somewhat better. In summary, HL-based SAFE-MIL achieved the lowest MSE and passed the HL test in both datasets, obtaining

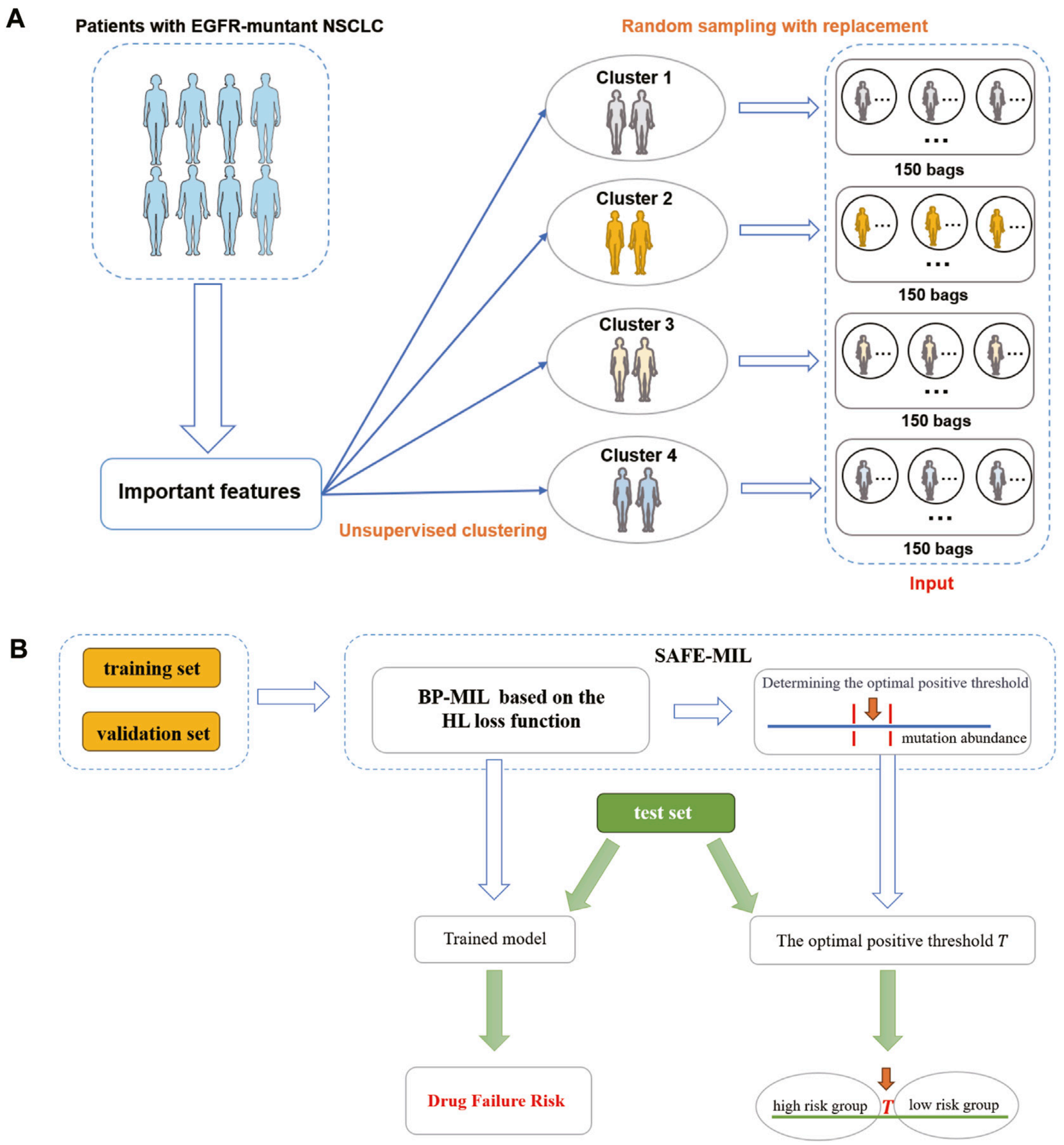

FIGURE 2
The flowchart of SAFE-MIL. **(A)** Data preprocessing process. First, cluster patients into 4 clusters using unsupervised learning. Then, randomly select 1–10 patients from each cluster to form a group, with each patient serving as an instance. **(B)** SAFE-MIL workflow. SAFE-MIL learns the representation of drug failure risk for each patient group through the HL loss function and determines the optimal positive threshold based on mutation abundance level.

TABLE 3 The input features of SAFE-MIL.

Feature	Data type	Description
Age	Numerical	The age of patients
Gender	Categorical	The gender of patients. A value of 1 corresponds to female, while 0 corresponds to male
Mutation abundance level	Numerical	Mutation Abundance level/Copy Number
TP53	Categorical	Does the patient have a TP53 gene mutation? A value of 1 indicates presence, while 0 indicates absence
The number of co-mutated genes	Numerical	The number of co-mutated genes

optimal HL statistics. Similarly, in the cohort2 and cohort3, SAFE-MIL based on the HL loss also achieved lower MSE (Figures 4, 5). Due to the larger sample size of cohort 2 compared to cohort1, theoretically, the model of HL-based SAFE-MIL predictive performance should be somewhat better. In summary, HL-based SAFE-MIL achieved the lowest

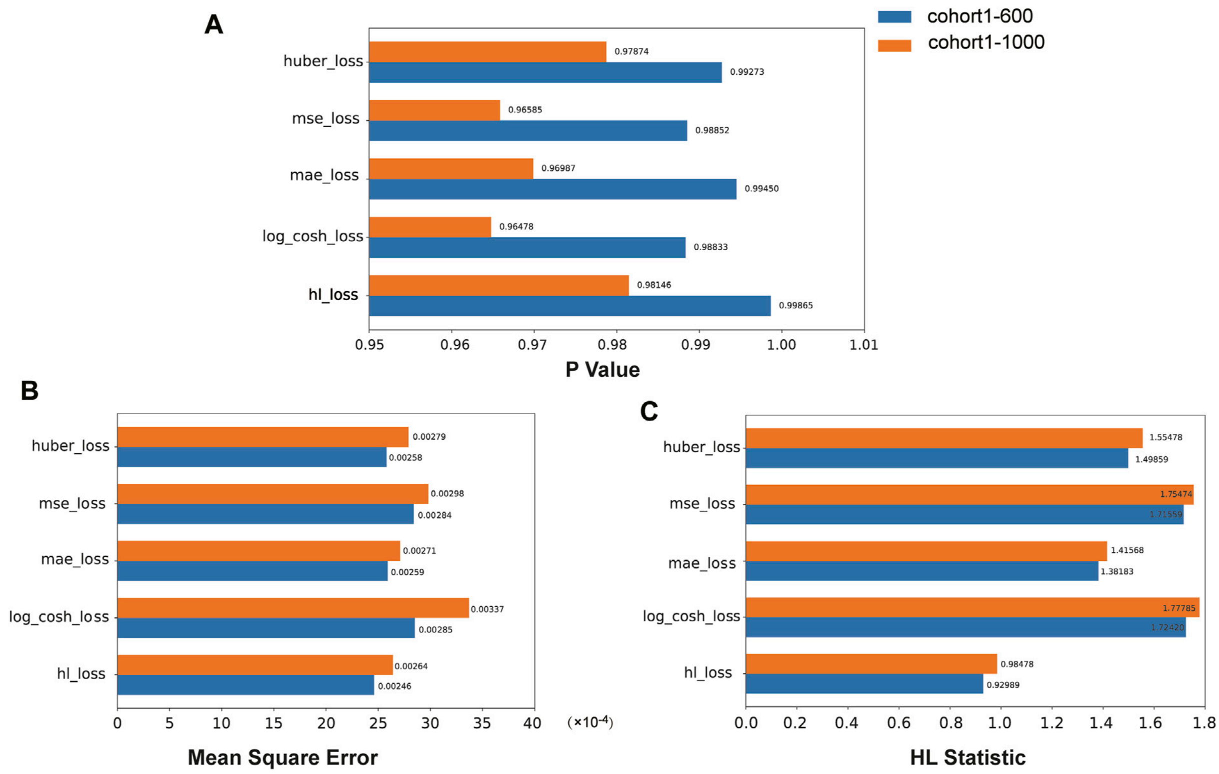

FIGURE 3
The performance of the HL loss function and four traditional regression loss functions were compared on cohort one in terms of *P*-value **(A)**, mean squared error **(B)**, and HL statistic **(C)**.

MSE and passed the HL test in all three datasets, obtaining optimal HL statistics.

Additionally, we selected cohort1-600 as examples to illustrate the relationship between the optimal positive threshold and drug efficacy. Based on the optimal positive threshold determined of mutation abundance, we divided patients in cohort1-600 into two groups to observe differences in drug failure risk and survival outcomes. Following the calculation rules mentioned earlier, the optimal positive threshold for mutation abundance in cohort1-600 was determined to be τ = 0.479. Firstly, using the threshold τ, we categorized patients into high-risk (mutation abundance level < τ) and low-risk (mutation abundance level >= τ) groups. In cohort1-600, the average drug failure risk in the high-risk group was 0.715, significantly higher than the 0.332 in the low-risk group (t-test, $p<$ 2.2e-16, Figure 6A). Since survival information for patients in cohort1-600 was recorded, we evaluated the survival differences between the high-risk and low-risk groups in cohort1-600. The results indicated that the average progression-free survival in the high-risk group in cohort1-600 was 8.4 months, significantly lower than the 16.4 months in the low-risk group (Log-rank test, p = 2.8e-9, Figure 6B).

The optimal positive threshold identified by SAFE-MIL represents the mutation abundance level at which patients are stratified into high-risk and low-risk groups with respect to treatment failure. This threshold is determined based on statistical analysis and represents a critical point for distinguishing patients who are more likely to experience treatment failure from those who are likely to respond favorably to therapy. Patients identified as high-risk above the threshold may benefit from closer monitoring, alternative treatment options, or enrollment in clinical trials for novel therapies. Conversely, patients below the threshold may be considered lower risk and may continue with standard therapy with confidence in treatment efficacy. The use of the optimal positive threshold enhances patient risk stratification by providing a quantitative measure that correlates with treatment outcomes. This allows for more accurate risk assessment and personalized treatment approaches, ultimately improving patient outcomes and optimizing resource allocation in healthcare settings. The identification of an optimal positive threshold enables clinicians to make informed decisions regarding treatment selection by considering individual patient risk profiles. SAFE-MIL facilitates precision medicine approaches, ensuring that patients receive the most appropriate and effective treatments based on their unique characteristics.

4 Discussion

The traditional clinical decision-making approaches often simplify drug efficacy into binary classifications, which fails to capture the nuanced risks associated with drug failure in individual patients. To address this limitation, we proposed that

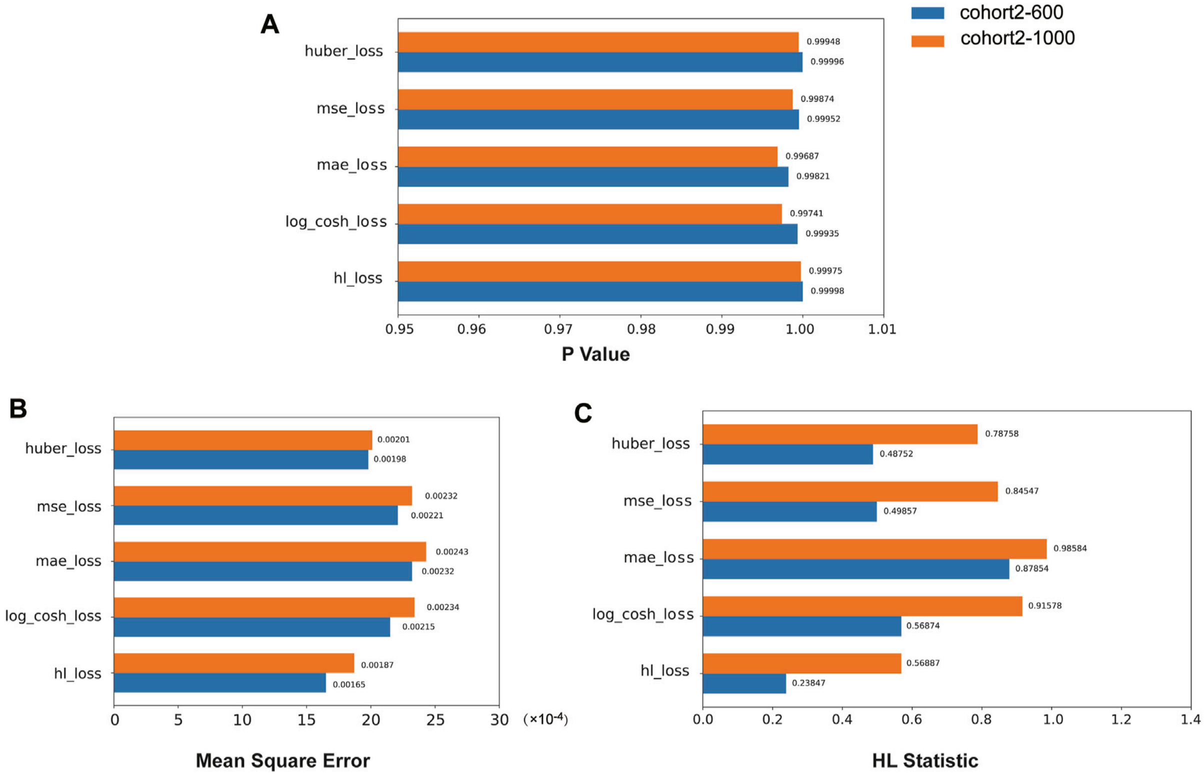

FIGURE 4
The performance of the HL loss function and four traditional regression loss functions were compared on cohort 2 in terms of *P*-value **(A)**, mean squared error **(B)**, and HL statistic **(C)**.

the clinical drug response for patients should be a regression value, reflecting individual variability in drug failure risk. This study introduces the SAFE-MIL framework, a new method for calculating and predicting the risk of drug failure in patients, which designs drug effectiveness labels based on clinical features of patients and builds a model to predict the risk of drug failure using multiple instance learning. A critical enhancement in SAFE-MIL is the integration of a loss function derived from the Hosmer-Lemeshow test. This integration not only boosts model performance but also ensures statistical interpretability, enabling robust assessments of drug failure risks and facilitating patient stratification by identifying the optimal positive threshold. Our results indicate that SAFE-MIL outperforms traditional regression methods in accuracy, particularly in capturing inter-patient variability in risk assessments, and successfully passes the HL test.

The concept of drug failure risk introduced in this study provides a more comprehensive characterization of drug effectiveness. Due to the significant molecular heterogeneity observed in tumors, there are often many different molecular features and feature combinations that can lead to the model predicting specific drug responses. However, elucidating these features and discerning whether they are distinct or functionally linked can be challenging. This is primarily due to the “black box” nature of most machine learning models, which prioritize predictive accuracy over an understanding of the underlying biological mechanisms (Ching et al., 2018). SAFE-MIL addresses this by enhancing clinical interpretability through its novel use of MIL and HL test integration, thereby somewhat bridging the gap between clinical needs and machine learning methodologies.

Furthermore, the study critically evaluates and expands upon existing methods for interpreting deep learning algorithms in clinical settings [(Zhang et al., 2018)]. Our approach incorporates the Hosmer-Lemeshow test into the MIL framework to estimate the risk of treatment failure, significantly improving the model’s predictive calibration performance and offering a unique risk assessment perspective beyond traditional interpretation methods. The results demonstrate that our framework outperforms traditional regression loss functions in terms of accuracy and successfully passes the Hosmer-Lemeshow test, highlighting its capability to accurately capture inter-patient variability in risk while ensuring statistical interpretability. Additionally, our framework enables the identification of an optimal threshold for mutation abundance level, allowing for effective patient stratification into high-risk and low-risk groups. This stratification reveals significant disparities in drug failure risk and progression-free survival, showcasing the utility of our approach in guiding treatment decisions.

Although our study primarily focuses on patients with EGFR-mutant non-small cell lung cancer treated with EGFR tyrosine kinase inhibitors, the potential of SAFE-MIL extends far beyond this specific context. Moving forward, efforts to validate and refine the SAFE-MIL framework across different cancer types and therapeutic modalities will further enhance its utility and broaden its impact on clinical practice. Beyond its predictive capabilities,

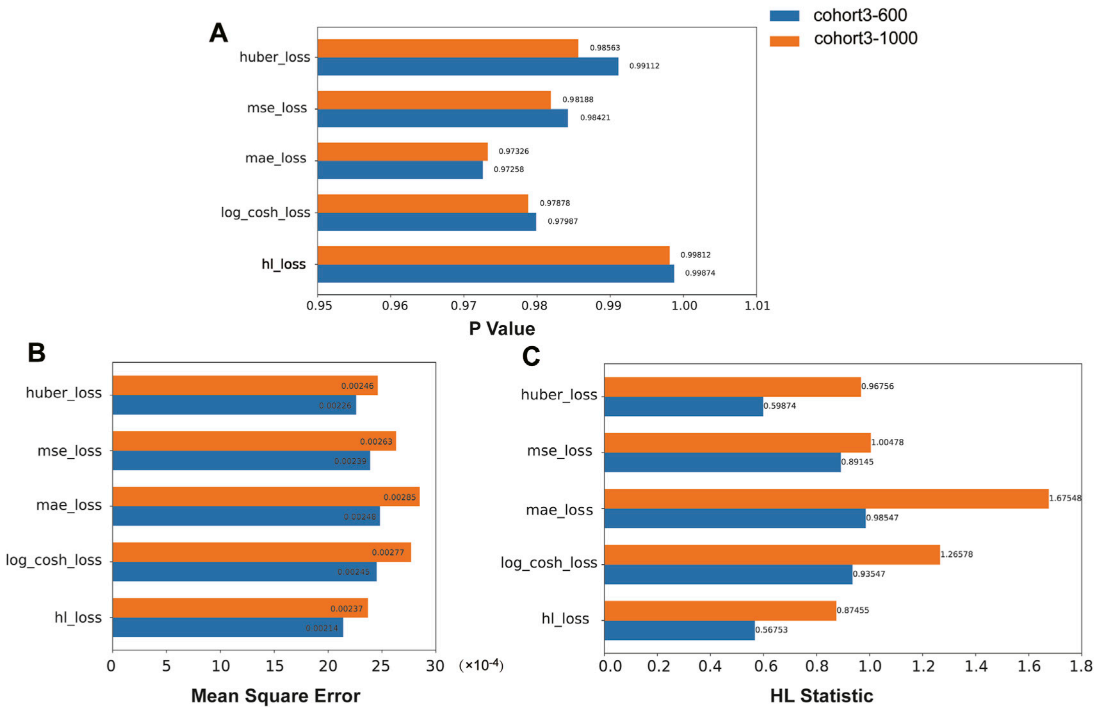

FIGURE 5
The performance of the HL loss function and four traditional regression loss functions were compared on cohort 3 in terms of *P*-value **(A)**, mean squared error **(B)**, and HL statistic **(C)**.

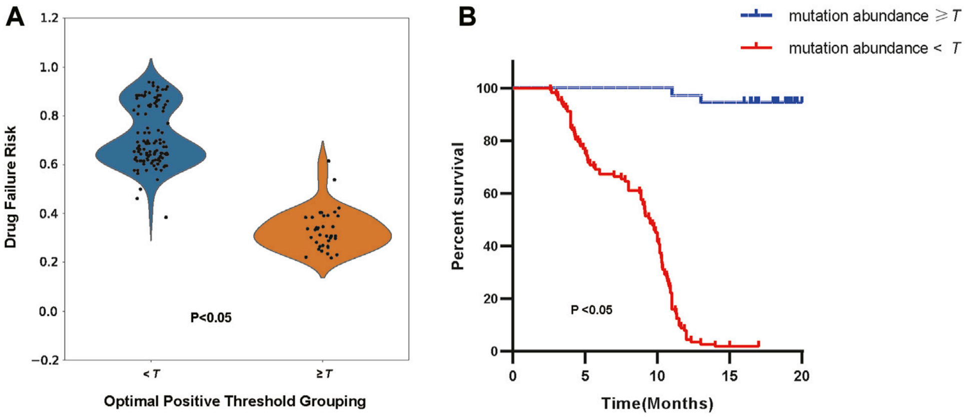

FIGURE 6
(A) The distribution of drug failure risk under the optimal positive threshold grouping (Independent Samples t-test, p < 2.2e-16). **(B)** The survival analysis of patients according to the optimal positive threshold (Log-rank test, p = 2.8e-9).

SAFE-MIL offers early risk identification and optimizing clinical trial design. For instance, SAFE-MIL enables the identification of patient subgroups with distinct risk profiles, paving the way for targeted interventions and precision medicine approaches. Moreover, integrating automated risk assessment tools into clinical workflows holds promise for streamlining decision-making processes and improving patient outcomes in real-world practice. Consequently, the integration of SAFE-MIL is poised to transform clinical practice by boosting treatment efficacy and ensuring patient safety. The drug failure risk metrics derived from SAFE-MIL appear to correlate with the patient's resistance mechanisms, a relationship that warrants deeper investigation.

Despite these promising results, our study acknowledges several limitations. Firstly, due to the limited number of patients, we could only generate groups through bootstrap sampling, and we constrained the number of groups and the number of instances per group. Additionally, the availability and quality of data sources pose significant challenges to model development and validation.

Future efforts should focus on employing SAFE-MIL in larger-scale studies to explore more robust and efficient risk calculation models. Furthermore, implementing SAFE-MIL in real-world healthcare settings necessitates careful consideration of practical challenges such as workflow integration, clinician training, and patient acceptance.

5 Conclusion

This paper presented SAFE-MIL, a novel risk assessment framework tailored to address the issue of varying survival benefits among patients with certain target gene mutations undergoing targeted therapy. By incorporating multiple instance learning (MIL), SAFE-MIL constructs patient-specific effectiveness labels and designs a novel interpretable loss function based on the Hosmer-Lemeshow test. This risk assessment framework accurately estimates the risk of treatment failure and also provides the optimal threshold for risk stratification. A comprehensive case study involving 457 non-small cell lung cancer patients with EGFR mutations treated with EGFR tyrosine kinase inhibitors demonstrated that SAFE-MIL outperforms conventional regression methods in accuracy, effectively capturing inter-patient variability in risk. This computational framework possesses statistical interpretability and adaptability, rendering it a helpful tool in clinical decision-making for targeted therapy. It has the potential to advance personalized medicine by enhancing patient stratification methods.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, and the data presented in the study are deposited in the GSA- human repository, accession number HRA006927.

Ethics statement

The studies involving humans were approved by the Institutional Review Board of Shanghai Chest Hospital (No. KS1707). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

YG: Data curation, Methodology, Software, Writing–original draft, Writing–review and editing. ZX: Formal Analysis, Methodology, Software, Writing–original draft, Writing–review and editing. JW: Conceptualization, Methodology, Supervision, Writing–review and editing. XA: Data curation, Resources, Writing–review and editing. RC: Resources, Supervision, Writing–review and editing. XY: Resources, Supervision, Writing–review and editing. SL: Conceptualization, Resources, Supervision, Writing–review and editing. YL: Conceptualization, Formal Analysis, Methodology, Supervision, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by National Natural Science Foundation of China, grant numbers 72274152, 82030045, 82241227, National Multi-disciplinary Treatment Project for Major Diseases, grant number 2020NMDTP, Collaborative Innovation Center for Clinical and Translational Science by Ministry of Education and Shanghai, grant numbers TM202112, CCTS202204, General program of Natural Science Foundation of Shanghai, grant number 22ZR1457500.

Acknowledgments

We thank all faculty members and graduate students who discussed the mathematical and statistical issues in seminars.

Conflict of interest

Authors YG, RC and XY were employed by GenePlus Beijing Institute.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1381851/full#supplementary-material

References

Amar, R. A., Dooly, D. R., Goldman, S. A., and Zhang, Q. (2001). Multiple-instance learning of real-valued data. *InICML* 28, 3–10. doi:10.5555/944919.944949

Banerjee, S., Mohammed, A., Wong, H. R., Palaniyar, N., and Kamaleswaran, R. (2021). Machine learning identifies complicated sepsis course and subsequent mortality based on 20 genes in peripheral blood immune cells at 24 H post-ICU admission. *Front. Immunol.* 12, 592303. doi:10.3389/fimmu.2021.592303

Blakely, C. M., Watkins, T. B. K., Wu, W., Gini, B., Chabon, J. J., McCoach, C. E., et al. (2017). Evolution and clinical impact of co-occurring genetic alterations in advanced-stage EGFR-mutant lung cancers. *Nat. Genet.* 49 (12), 1693–1704. doi:10.1038/ng.3990

Blumenthal, G. M., Gong, Y., Kehl, K., Mishra-Kalyani, P., Goldberg, K. B., Khozin, S., et al. (2019). Analysis of time-to-treatment discontinuation of targeted therapy, immunotherapy, and chemotherapy in clinical trials of patients with non-small-cell lung cancer. *Ann. Oncol.* 30 (5), 830–838. doi:10.1093/annonc/mdz060

Chang, L., Wu, J., Moustafa, N., Bashir, A. K., and Yu, K. (2022). AI-driven synthetic biology for non-small cell lung cancer drug effectiveness-cost analysis in intelligent assisted medical systems. *IEEE J. Biomed. Health Inf.* 26 (10), 5055–5066. doi:10.1109/JBHI.2021.3133455

Chen, Z., Wei, S., Yuan, Z., Chang, R., Chen, X., Fu, Y., et al. (2024). Machine learning reveals ferroptosis features and a novel ferroptosis classifier in patients with sepsis. *Immun. Inflamm. Dis.* 12 (5), e1279. doi:10.1002/iid3.1279

Cheng, Y., Ma, L., Liu, Y., Zhu, J., Xin, Y., Liu, X., et al. (2020). Comprehensive characterization and clinical impact of concomitant genomic alterations in EGFR-mutant NSCLCs treated with EGFR kinase inhibitors. *Lung Cancer* 145, 63–70. doi:10.1016/j.lungcan.2020.04.004

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* 15 (141), 20170387. doi:10.1098/rsif.2017.0387

Daoud, S., Mdhaffar, A., Jmaiel, M., and Freisleben, B. (2020). Q-rank: reinforcement learning for recommending algorithms to predict drug sensitivity to cancer therapy. *IEEE J. Biomed. Health Inf.* 24 (11), 3154–3161. doi:10.1109/JBHI.2020.3004663

Davies, D. R., Redfield, M. M., Scott, C. G., Minamisawa, M., Grogan, M., Dispenzieri, A., et al. (2022). A simple score to identify increased risk of transthyretin amyloid cardiomyopathy in heart failure with preserved ejection fraction. *JAMA Cardiol.* 7 (10), 1036–1044. doi:10.1001/jamacardio.2022.1781

Diao, Y., Zhao, Y., Li, X., Li, B., Huo, R., and Han, X. (2023). A simplified machine learning model utilizing platelet-related genes for predicting poor prognosis in sepsis. *Front. Immunol.* 14, 1286203. doi:10.3389/fimmu.2023.1286203

Fisher, R. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* 10 (4), 507–521. doi:10.2307/2331838

Fisher, R. (1921). On the 'probable error' of a coefficient of correlation deduced from a small sample. *Contributions Math. Statistics*, 3–32.

Fu, G., Nan, X., Liu, H., Patel, R. Y., Daga, P. R., Chen, Y., et al. (2012). Implementation of multiple-instance learning in drug activity prediction. *BMC Bioinforma.* 13 (15), S3. doi:10.1186/1471-2105-13-S15-S3

Göttlich, C., Müller, L. C., Kunz, M., Schmitt, F., Walles, H., Walles, T., et al. (2016). A combined 3D tissue engineered *in vitro*/*in silico* lung tumor model for predicting drug effectiveness in specific mutational backgrounds. *J. Vis. Exp.* 110, e53885. doi:10.3791/53885

Hosmer, D. W., and Lemesbow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Commun. Statistics - Theory Methods* 9 (10), 1043–1069. doi:10.1080/03610928008827941

Huber, P. J. (1992). Robust estimation of a location parameter. *Break. statistics Methodol. distribution*, 492–518. doi:10.1007/978-1-4612-4380-9_35

Kramer, A. A., and Zimmerman, J. E. (2007). Assessing the calibration of mortality benchmarks in critical care: the Hosmer-Lemeshow test revisited. *Crit. Care Med.* 35 (9), 2052–2056. doi:10.1097/01.CCM.0000275267.64078.B0

Kuenzi, B. M., Park, J., Fong, S. H., Sanchez, K. S., Lee, J., Kreisberg, J. F., et al. (2020). Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell* 38 (5), 672–684.e6. doi:10.1016/j.ccell.2020.09.014

Liu, Y., Wang, H., Yang, S., Yang, Y., Wu, Y., He, Z., et al. (2022). EGFR mutation types and abundance were associated with the overall survival of advanced lung adenocarcinoma patients receiving first-line tyrosine kinase inhibitors. *J. Thorac. Dis.* 14 (6), 2254–2267. doi:10.21037/jtd-22-755

Łosińska, K., Wilk, M., Pripp, A. H., Korkosz, M., and Haugeberg, G. (2022). Long-term drug effectiveness and survival for reference rituximab in rheumatoid arthritis patients in an ordinary outpatient clinic. *Sci. Rep.* 12 (1), 8283. doi:10.1038/s41598-022-12271-9

Nemati, S., Holder, A., Razmi, F., Stanley, M. D., Clifford, G. D., and Buchman, T. G. (2018). An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit. care Med.* 46 (4), 547–553. doi:10.1097/CCM.0000000000002936

Nong, J., Gong, Y., Guan, Y., Yi, X., Yi, Y., Chang, L., et al. (2018). Circulating tumor DNA analysis depicts subclonal architecture and genomic evolution of small cell lung cancer [published correction appears in Nat Commun 2019 Jan 29;10(1):552]. *Nat. Commun.* 9 (1), 3114. doi:10.1038/s41467-018-05327-w

Paz-Ares, L., Luft, A., Vicente, D., Tafreshi, A., Gümüş, M., Mazières, J., et al. (2018). Pembrolizumab plus chemotherapy for squamous non-small-cell lung cancer. *N. Engl. J. Med.* 379 (21), 2040–2051. doi:10.1056/NEJMoa1810865

Peng, W., Chen, T., and Dai, W. (2022). Predicting drug response based on multi-omics fusion and graph convolution. *IEEE J. Biomed. Health Inf.* 26 (3), 1384–1393. doi:10.1109/JBHI.2021.3102186

Robichaux, J. P., Le, X., Vijayan, R. S. K., Hicks, J. K., Heeke, S., Elamin, Y. Y., et al. (2021). Structure-based classification predicts drug response in EGFR-mutant NSCLC. *Nature* 597 (7878), 732–737. doi:10.1038/s41586-021-03898-1

Rubin, E. H., and Gilliland, D. G. (2012). Drug development and clinical trials--the path to an approved cancer drug. *Nat. Rev. Clin. Oncol.* 9 (4), 215–222. doi:10.1038/nrclinonc.2012.22

Saberian, M. S., Moriarty, K. P., Olmstead, A. D., Hallgrimson, C., Jean, F., Nabi, I. R., et al. (2022). DEEMD: drug efficacy estimation against SARS-CoV-2 based on cell morphology with deep multiple instance learning. *IEEE Trans. Med. Imaging* 41 (11), 3128–3145. doi:10.1109/TMI.2022.3178523

Samstein, R. M., Lee, C. H., Shoushtari, A. N., Hellmann, M. D., Shen, R., Janjigian, Y. Y., et al. (2019). Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat. Genet.* 51 (2), 202–206. doi:10.1038/s41588-018-0312-8

Schnipper, L. E., Davidson, N. E., Wollins, D. S., Tyne, C., Blayney, D. W., Blum, D., et al. (2015). American society of clinical oncology statement: a conceptual framework to assess the value of cancer treatment options. *J. Clin. Oncol.* 33 (23), 2563–2577. doi:10.1200/JCO.2015.61.6706

Shen, X., Tian, X., Liu, T., Xu, F., and Tao, D. (2018). Continuous dropout. *IEEE Trans. Neural Netw. Learn Syst.* 29 (9), 3926–3937. doi:10.1109/TNNLS.2017.2750679

Sotudian, S., and Paschalidis, I. C. (2022). Machine learning for pharmacogenomics and personalized medicine: a ranking model for drug sensitivity prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform* 19 (4), 2324–2333. doi:10.1109/TCBB.2021.3084562

Tang, X., Chen, X., Liao, Y., Yan, B., Hu, H., Ming, Z., et al. (2021). Self-internal-reference probe system for control-free quantification of mutation abundance. *Anal. Chem.* 93 (39), 13274–13283. doi:10.1021/acs.analchem.1c02877

Wang, C., Chen, Y., Zhao, L., Wang, J., and Wen, N. (2022). Modeling DTA by combining multiple-instance learning with a private-public mechanism. *Int. J. Mol. Sci.* 23 (19), 11136. doi:10.3390/ijms231911136

Wang, X., Liu, Y., Meng, Z., Wu, Y., Wang, S., Jin, G., et al. (2021). Plasma EGFR mutation abundance affects clinical response to first-line EGFR-TKIs in patients with advanced non-small cell lung cancer. *Ann. Transl. Med.* 9 (8), 635. doi:10.21037/atm-20-7155

Yan, X., Wang, H., Li, P., Zhang, G., Zhang, M., Yang, J., et al. (2019). Efficacy of first-line treatment with epidermal growth factor receptor-tyrosine kinase inhibitor (EGFR-TKI) alone or in combination with chemotherapy for advanced non-small cell lung cancer (NSCLC) with low-abundance mutation. *Lung Cancer* 128, 6–12. doi:10.1016/j.lungcan.2018.12.007

Yang, K., Yang, Y., Fan, S., Xia, J., Zheng, Q., Dong, X., et al. (2023). DRONet: effectiveness-driven drug repositioning framework using network embedding and ranking learning. *Brief. Bioinform* 24 (1), bbac518. doi:10.1093/bib/bbac518

Zhang, L., Zhang, F., Xu, F., Wang, Z., Ren, Y., Han, D., et al. (2021b). Construction and evaluation of a sepsis risk prediction model for urinary tract infection. *Front. Med. (Lausanne)* 8, 671184. doi:10.3389/fmed.2021.671184

Zhang, Y., Chang, L., Yang, Y., Fang, W., Guan, Y., Wu, A., et al. (2019). The correlations of tumor mutational burden among single-region tissue, multi-region tissues and blood in non-small cell lung cancer. *J. Immunother. Cancer* 7 (1), 98. doi:10.1186/s40425-019-0581-5

Zhang, Y., Hu, X., Li, Y., Zhao, C-X., Abudushalamu, Jin, X-L., et al. (2021a). International clinical practice guideline of Chinese medicine Alzheimer. *World J. Trad. Chin. Med.* 7 (2), 265–275. doi:10.4103/wjtcm.wjtcm_28_21

Zhang, Z., Beck, M. W., Winkler, D. A., Huang, B., Sibanda, W., Goyal, H., et al. (2018). Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Ann. Transl. Med.* 6 (11), 216. doi:10.21037/atm.2018.05.32

Zhang, Z., Pan, Q., Ge, H., Xing, L., Hong, Y., and Chen, P. (2020). Deep learning-based clustering robustly identified two classes of sepsis with both prognostic and predictive values. *EBioMedicine* 62, 103081. doi:10.1016/j.ebiom.2020.103081

Zhao, Z., Fu, G., Liu, S., Elokely, K. M., Doerksen, R. J., Chen, Y., et al. (2013). Drug activity prediction using multiple-instance learning via joint instance and feature selection. *BMC Bioinforma.* 14 (Suppl. 14), S16. doi:10.1186/1471-2105-14-S14-S16

Zhou, Q., Zhang, X. C., Chen, Z. H., Yin, X. L., Yang, J. J., Xu, C. R., et al. (2011). Relative abundance of EGFR mutations predicts benefit from gefitinib treatment for advanced non-small-cell lung cancer. *J. Clin. Oncol.* 29 (24), 3316–3321. doi:10.1200/JCO.2010.33.3757