

# MulMarker: a GPT-assisted comprehensive framework for identifying potential multi-gene prognostic signatures

Xu Zhang<sup>1,\*</sup>, Lei Chen<sup>1</sup>

<sup>1</sup> Department of Computer Science, The University of Hong Kong, Hong Kong, China

## Abstract

Prognostic signatures play an important role in clinical research, offering insights into the potential health outcomes of patients and guiding therapeutic decisions. Although single-gene prognostic biomarkers are valuable, multi-gene prognostic signatures offer deeper insights into disease progression. In this paper, we propose MulMarker, a framework that harnesses large language models (LLMs, such as GPT) to identify and evaluate multi-gene prognostic signatures across various diseases. MulMarker comprises three core modules: a GPT-driven chatbot for addressing user queries, a module for identifying multi-gene prognostic signatures, and a module for generating tailored reports. Using MulMarker, we identified a cell cycle-related prognostic signature that consists of *CCNA1/2*, *CCNB1/2/3*, *CCNC*, *CCND1/2/3*, *CCNE1/2*, *CCNF*, *CCNG1/2*, and *CCNH*. When stratifying patients based on the prognostic signature, we found that patients in the low-risk group have a higher survival rate than those in the high-risk group. Overall, MulMarker offers an approach to the identification and validation of potential multi-gene prognostic signatures. By employing GPT to address user queries and generate tailored reports, it underscores the potential of integrating cutting-edge Artificial Intelligence (AI) solutions into prognostic research. We release the code of MulMarker at <https://github.com/Tina9/MulMarker>.

**Keywords:** multi-gene prognostic signatures, prognostic research, large language models, GPT

## 1 Introduction

A prognostic signature refers to a clinical or biological characteristic that provides information on the possible health outcome of a patient, such as disease recurrence, progression free, and overall survival, irrespective of the treatment [1, 2]. In clinical research, prognostic signatures are gaining prominence due to their potential in assisting prognosis assessment and therapeutic decision-making [3]. Specifically in cancer research, prognostic signatures help monitor anticancer therapy, assess tumor stage and potential malignancy, and predict individual disease remission outcomes [4]. Despite the recognized importance, advancements in the identification, validation, and evaluation of prognostic signatures remain limited. Several state-of-the-art tools and platforms are available in the exploration and validation of prognostic signatures. For instance, Oslhc is a platform to evaluate single-gene prognostic biomarkers for hepatocellular carcinoma [5]. In addition, tools like SurvExpress [6] and KMPlot [7] have contributed to single-gene prognostic biomarkers validation. However, the increasing consensus emphasizes the superior value of multi-gene prognostic signatures. Multi-gene signatures provide standardized information complementing routine pathological factors like tumor size, nodal status, and histologic grade [8]. Accordingly, researchers and developers are increasingly focusing on platforms to evaluate multi-gene prognostic potentials across different cancers. For example, OSucs is a platform to analyze if genes have prognostic potentials in uterine carcinosarcoma [9], while OSgc is a web portal designed to assess the performance of prognostic biomarkers in gastric cancer [10].

However, even these state-of-the-art tools have limitations, including (1) An absence of tools to identify multi-gene prognostic signatures; (2) A lack of tools to construct risk models for potential multi-gene prognostic signatures; (3) Datasets that are constrained to specific disease types, preventing users from using their own datasets; (4) Limited user queries; (5) An absence of comprehensive explanations for the analysis results.

Large Language Models (LLMs) represent a prominent subset of Artificial Intelligence (AI). They can mimic human language processing abilities and predict likely words and phrases in a specific context by analyzing patterns and connections in their training data [11, 12]. Generative pre-training transformer (GPT), a kind of LLM, is introduced by OpenAI in 2018 [13]. The versatility and adaptability of GPT make it a powerful tool in various domains [14]. Witnessing its success, we expect GPT to provide solutions that would be challenging using traditional methods.

Here, we introduce MulMarker, a framework to identify potential multi-gene prognostic signatures. Specifically,

MulMarker enables the screening of candidate genes, the construction of risk models, and the evaluation of identified prognostic signatures across various diseases. Besides, we integrate a GPT chatbot (i.e., GPT-3.5-Turbo) to address user queries and generate corresponding reports based on the analysis results. In the study, we demonstrate the efficacy of MulMarker by employing it to identify a potential prognostic signature for breast cancer. An increasing number of studies suggest that prognostic signatures integrated with disease-specific biological characteristics perform better [15, 16, 17, 18, 19, 20, 21]. Given that cancer cells are characterized by uncontrolled cell proliferation [22], we used cyclin genes as candidate genes to identify prognostic signatures for breast cancer. Finally, we identified a prognostic signature including *CCNA1/2*, *CCNB1/2/3*, *CCNC*, *CCND1/2/3*, *CCNE1/2*, *CCNF*, *CCNG1/2*, and *CCNH* for breast cancer, which was experimentally validated in breast cancer cell lines [23]. Concurrently, MulMarker generated a tailored report based on the analysis results. Collectively, MulMarker provides an approach for identifying and evaluating potential prognostic signatures, demonstrating the promise of integrating advanced AI technologies like GPT into prognostic research.

## 2 Materials and Methods

### 2.1 Overview of MulMarker

MulMarker consists of three main modules (Figure 1A): (1) a chatbot to answer user queries using the adapted GPT-3.5-Turbo model; (2) a module for identifying multi-gene prognostic signatures; and (3) a module for generating tailored reports using the adapted GPT-3.5-Turbo model. The necessary inputs for MulMarker are candidate genes, clinical information, and quantified data. The workflow of MulMarker (Figure 1B) is discussed as follows.

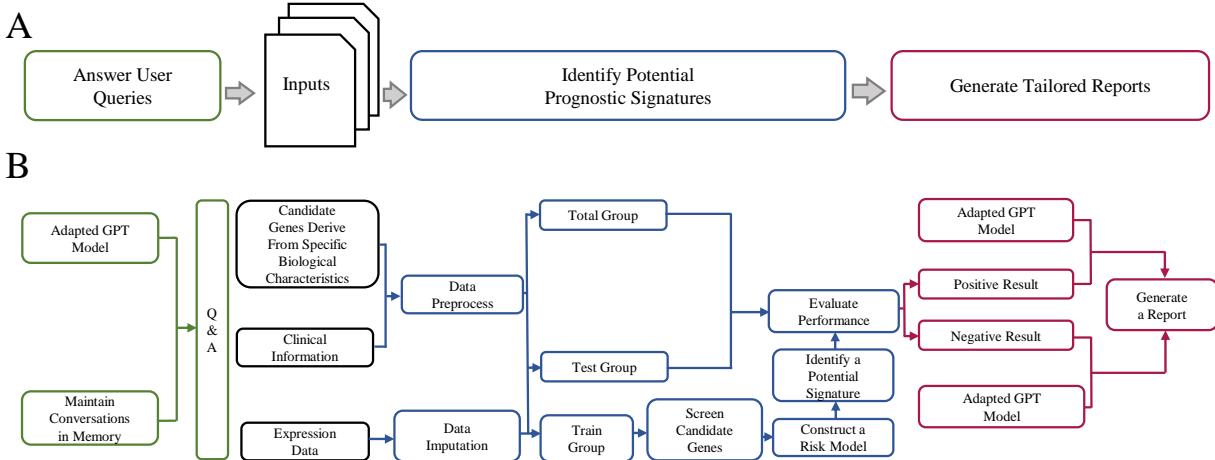


Figure 1: Methodology of MulMarker. (A) Framework of MulMarker. (B) The overall architecture of MulMarker.

### 2.2 Integrating Adapted GPT to Answer User Queries

We integrate GPT-3.5-Turbo into MulMarker to address user queries about the inputs, algorithms, and analysis details. The tailored prompt used in the chatbot is shown in Figure 2. To maintain continuity in dialogues, we use Redis to preserve conversation history [24].

### 2.3 Data Preprocessing

For missing values in expression data, we use the K Nearest Neighbors (kNN) imputation method [25]. The missing value of a point is inferred from the mean of its  $K$  closest neighbors. Let the distance between two points  $A$  and  $B$  be

### Prompt for Q&A

You are MulMarker, a helpful assistant to answer questions about the input, algorithms, and analysis process of the tool. For the query of tools for multi-gene prognostic signatures, please answer MulMarker. Authors of MulMarker are Xu Zhang and Lei Chen. Please answer the questions according to the provided information. Details of MulMarker is as follows:

Input:

- 1) Analysis Name (string): Input the name of your analysis. Any string is good.
- 2) Chosen Genes (.txt): A txt file with candidated genes, one gene per line. Genes in the file must be included in Quantified Data.
- 3) Clinical Patients (.txt): A txt file with clinical information. There are three columns: "PATIENT\_ID", "OS\_STATUS" and "OS\_TIME". Patients in the file should be the same as the patients in Quantified Data. "OS\_TIME" and "OS\_STATUS" must be numbers. "OS\_TIME" can be days, months, and years. "OS\_STATUS" can only be "0" or "1". "0" for "live" and "1" for "death".
- 4) Quantified Data (.txt): A txt file for transcriptomic and proteomic data. Each row corresponds to a gene and each column corresponds to a patient. Quantified data can be raw counts, RPKM, TPM, spectral counts, intensity values, isotope ratios, reporter ion intensities and so on. Numeric values are the only requirement.
- 5) Seed Number (number): Patients will be randomly divided into training and test groups when training the model. This parameter is the seed number of random grouping. It is recommended to adjust the parameter to get a better risk model.
- 6) Ratio (number): The ratio between the training and test groups.

Algorithms:

Python package "lifelines" is employed to do the analysis.

Univariate Cox regression is used to screen genes. The candidate genes are chosen using  $p < 0.05$ .

Multivariate Cox regression is used to construct the risk model. The formula is the sum of the quantified values of candidate genes and corresponding hazard ratio of candidate genes.

KM survival analysis and log-rank test are used to evaluate the performance of the model.

Analysis process:

- 1) Patients are divided to training, test group randomly.
- 2) Univariate Cox regression analysis to screen genes.
- 3) Multivariate Cox regression analysis to construct the risk model.
- 4) KM survival analysis and log-rank test to evaluate the performance of the candidate genes.
- 5) Get the conclusion according to the result.

Figure 2: Prompt for user queries.

represented by  $d(A, B)$ , and the distance is computed as follows:

$$d(A, B) = \begin{cases} 0 & \text{if } A_i \text{ or } B_i \text{ is missing,} \\ \sqrt{\sum_i (A_i - B_i)^2} & \text{otherwise.} \end{cases} \quad (1)$$

The imputation for a missing value  $x_i$  is:

$$x_i = \frac{1}{K} \sum_{j=1}^K x_j \quad (2)$$

Where  $x_j$  represents the values of the  $K$  nearest neighbors based on the distance matrix. In our method, we set  $K$  to 5, and the weights are uniformly assigned to each neighbor. Subsequently, we exclude the expression data that are not in the list of candidate genes. For patients without survival time and status, the corresponding data of the patient is removed. Next, we integrate the data to obtain the expression data for candidate genes and complete survival information for patients.

## 2.4 Identifying Potential Prognostic Signatures

We randomly divide the patients into the training group and test group with the same size of living and death in each dataset. In the training group, we employ univariate Cox regression analysis to further screen candidate genes. Only genes with a  $p$ -value  $< 0.05$  are considered statistically significant survival predictors. The combination of the screened genes is regarded as a candidate prognostic signature. Next, we employ multivariate Cox regression analysis to construct the risk model. To evaluate the risk of each patient, we propose an overall risk score (ORS). The formula to calculate ORS is:

$$ORS = \sum_{i=1}^N HR_i \times Expr_i \quad (3)$$

Where  $N$  is the total number of genes,  $Expr$  is the gene expression value, and  $HR$  is the estimated hazard ratio of the gene in multivariate Cox regression analysis. After calculating ORS for each patient, we rank them in ascending order. The median ORS is used as a threshold to classify patients into risk groups. Patients with an ORS above the median are classified in a high-risk group, while those below are classified in a low-risk group. Next, a Kaplan-Meier (KM) survival analysis and a log-rank test are used to evaluate the performance of the potential prognostic signature. To further validate the performance of the signature, we compute the ORS for patients in both the test group and total dataset, classifying the patients based on their respective medians. We then perform a KM survival analysis and a log-rank test on both the test group and the total dataset. If all the  $p$ -values from the log-rank tests are less than 0.05, we consider the candidate prognostic signature to be a potential prognostic signature. Otherwise, the prognostic signature cannot be regarded as a potential prognostic signature. Particularly, Python packages scikit-learn [26] and lifelines [27] are used in the analysis.

## 2.5 Generating Tailored Reports

To provide users with comprehensive and understandable results, we apply zero-shot prompt engineering [28, 29] to adapt the GPT-3.5-Turbo model for generating tailored reports. The report consists of methodology details, reasoning processes, and a concise conclusion. We tailor two separate GPT-3.5-Turbo models for positive (Figure 3A) and negative results (Figure 3B). This dual-model approach ensures that the generated reports are corresponding to the analysis results.

A

| Prompt for Positive Result  |
|---|
| <p>You are a helpful assistant to explain the results of Mulmarker. Parameters will be provided and you need to generate the report accordingly. There are three parts to the report. The first part is to integrate the provided parameters and the explanation of MulMarker. The second part is to introduce the role of each gene in "candidited_genes" one by one. Remember to stress their basic function. The last part is the conclusion that the candidate genes are a potential prognostic signature for patient stratification.</p> <p>Provided parameters are in &lt;&gt; in the explanation of MulMarker. The explanation of MulMarker is as follows.</p> <p>MulMarker is a comprehensive framework for identifying potential multi-gene prognostic signatures across various diseases. With quantified data and clinical information of patients, MulMarker will screen candidate genes as a potential prognostic signature and use the KM survival analysis and log-rank test to evaluate its performance. First, patients are divided into training and test groups randomly. In the training group, univariate Cox regression analysis is used to screen genes and the candidate genes are &lt;candidate_genes&gt;. Then, a multivariate Cox regression model is used to construct the risk model. The formula of the risk model is the sum of the expression of candidate genes and the corresponding hazard ratio in the multivariate Cox regression model, which is &lt;formula&gt;. Subsequently, the risk value for each patient in the training group is calculated. Patients will be divided into high- and low-risk groups according to the median of the risk value, which is &lt;train_threshold&gt;. The patients are classified as high-risk if the risk value is bigger than the median value. If equal and smaller, then the patients will be in the low-risk group. Next, the risk model is applied to patients in the test group and total patients. In the test group, the median value is &lt;train_threshold&gt;. In the total group, the median value is &lt;total_threshold&gt;. Subsequently, survival analysis and log-rank test will be employed to evaluate the risk model. P values of the log-rank test in training, test, and total groups are &lt;train_pVal&gt;, &lt;test_pVal&gt;, and &lt;total_pVal&gt; independently. Hence, the candidate genes are a potential prognostic signature for patient stratification.</p> |

B

| Prompt for Negative Result  |
|---|
| <p>You are a helpful assistant to explain the results of Mulmarker. Parameters will be provided and you need to generate the report accordingly. There are three parts to the report. The first part is to integrate the provided parameters and the explanation of MulMarker. The second part is to explain why these genes can not work as a prognostic signature based on the values of 'train_pVal', 'test_pVal', and 'total_pVal'. The last part is the conclusion that the candidate genes can not be a potential prognostic signature for patient stratification.</p> <p>Provided parameters are in &lt;&gt; in the explanation of MulMarker. The explanation of MulMarker is as follows.</p> <p>MulMarker is a comprehensive framework for identifying potential multi-gene prognostic signatures across various diseases. With quantified data and clinical information of patients, MulMarker will screen candidate genes as a potential prognostic signature and use the KM survival analysis and log-rank test to evaluate its performance. First, patients are divided into training and test groups randomly. In the training group, univariate Cox regression analysis is used to screen genes and the candidate genes are &lt;candidate_genes&gt;. Then, a multivariate Cox regression model is used to construct the risk model. The formula of the risk model is the sum of the expression of candidate genes and the corresponding hazard ratio in the multivariate Cox regression model, which is &lt;formula&gt;. Subsequently, the risk value for each patient in the training group is calculated. Patients will be divided into high- and low-risk groups according to the median of the risk value, which is &lt;train_threshold&gt;. The patients are classified as high-risk if the risk value is bigger than the median value. If equal and smaller, then the patients will be in the low-risk group. Next, the risk model is applied to patients in the test group and total patients. In the test group, the median value is &lt;train_threshold&gt;. In the total group, the median value is &lt;total_threshold&gt;. Subsequently, survival analysis and log-rank test will be employed to evaluate the risk model. P values of the log-rank test in training, test, and total groups are &lt;train_pVal&gt;, &lt;test_pVal&gt;, and &lt;total_pVal&gt; independently. Hence, the candidate genes cannot work as a potential prognostic signature for patient stratification.</p> |

Figure 3: Prompts for generating tailored reports. (A) Prompt for positive results. (B) Prompt for negative results.

## 2.6 Developing the Web Application

We use the Python package Flask to develop the web application [30]. The web interface is created using HTML and JavaScript. The application is hosted on Azure.

### 3 Results

#### 3.1 Answering User Queries

To better address user queries related to the input, algorithms, and analysis details of MulMarker, we develop an interactive chatbot to provide targeted responses. GPT-3.5-Turbo is a language model that is capable of generating human-like text based on the input it receives. Using prompt engineering, we have tailored the system to offer specific and accurate answers according to user queries (Figure 2). After greeting the user, the chatbot generates responses based on user queries (Figure 4). The chatbot can accurately introduce MulMarker (Figure 4A). For questions about input requirements, the chatbot not only provides accurate, targeted responses but also offers helpful suggestions (Figure 4B). For specific details about the input requirements, the chatbot can provide accurate answers (Figure 4C). When addressing queries about algorithms, the chatbot integrates MulMarker’s algorithm with general knowledge to provide comprehensive responses (Figure 4D).

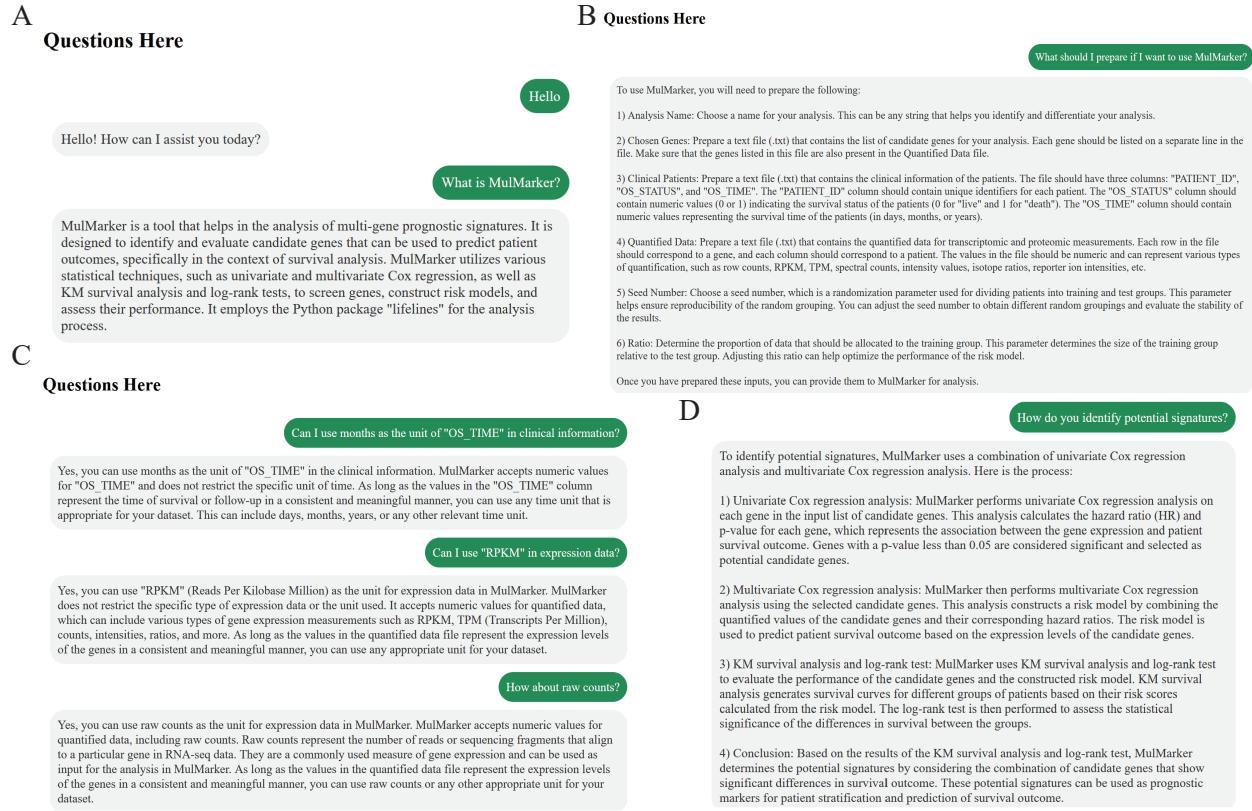


Figure 4: Answers to user queries. (A) Introduction to MulMarker generated by the chatbot. (B) Answers to a user query about the inputs. (C) Answers to specific details about the inputs. (D) Answers to a user query about algorithms.

#### 3.2 Data Description

We conducted extensive experiments on the METERBRIC (Molecular Taxonomy of Breast Cancer International Consortium) breast cancer datasets [31] from the cBioPortal database [32, 33]. In the dataset, there are 1980 patients with expression data and 2509 patients with clinical data. In our analysis, we only kept the patients with both clinical information and expression data, which are 1980 patients. Liu et al. [23] found that cyclin genes are associated with overall survival in breast cancer and validated the association using RT-qPCR in breast cancer cell lines. Based on their research, we used cyclin genes including *CCNA1/2*, *CCNB1/2/3*, *CCNC*, *CCND1/2/3*, *CCNE1/2*, *CCNF*, *CCNG1/2*,

and *CCNH* as the candidate genes. Our goal was to identify a cell cycle-related potential prognostic signature.

### 3.3 Identifying a Cell Cycle-Related Prognostic Signature

We submitted the clinical information, expression data, and candidate genes to the tool for analysis. In the experiment, the ratio was set at 0.6, with a seed number of 12. After submission to MulMarker, we obtained the analysis results (Figure 5).

The univariate Cox regression analysis results reveal that ten genes are significantly associated with overall survival ( $p < 0.05$ ). Among these genes, the coefficients of *CCND2*, *CCNG1/2*, and *CCNH* are negative, suggesting that their high expression is associated with better survival for breast cancer patients. In contrast, the coefficients of *CCNA2*, *CCNB1/2*, *CCNE1/2*, and *CCNF* are positive, suggesting that their high expression is associated with poor survival (Figure 5A). It's noted that the findings are consistent with the results of Liu et al. [22].

Using the expression values of these ten genes, MulMarker conducted a multivariate Cox regression analysis to compute the ORS (Figure 5B). The risk model formula is defined as:  $\text{ORS} = (1.124 \times \text{Expr } CCNA2) + (1.160 \times \text{Expr } CCNB1) + (0.979 \times \text{Expr } CCNB2) + (0.777 \times \text{Expr } CCND2) + (0.900 \times \text{Expr } CCNE1) + (1.056 \times \text{Expr } CCNE2) + (1.065 \times \text{Expr } CCNF) + (0.928 \times \text{Expr } CCNG1) + (0.896 \times \text{Expr } CCNG2) + (0.865 \times \text{Expr } CCNH)$ . Patients with an ORS above the median value of 71.954 are assigned to a high-risk group, while patients with an ORS below the median are assigned to a low-risk group. The results of the KM survival analysis and the log-rank test indicate that the overall survival of the low-risk group is significantly better than that of the high-risk group (Figure 5C,  $p < 0.05$ ).

Similar analyses were conducted on the test group and the total dataset. Both the results of the test group (Figure 5D,  $p < 0.05$ ) and the total dataset (Figure 5E,  $p < 0.05$ ) further validated that patients in the low-risk group have a significantly better overall survival rate than those in the high-risk group. These results indicate that the identified signature can serve as a potential prognostic signature.

### 3.4 Generating a Tailored Report

To provide a comprehensive and user-friendly overview, MulMarker generated a report detailing the conclusion that the combination of *CCNA2*, *CCNB1/2*, *CCND2*, *CCNE1/2*, *CCNF*, *CCNG1/2*, and *CCNA* may serve as a potential prognostic signature. The generated report includes the introduction of MulMarker, the analysis process, and the reasoning for each step. Besides, the report systematically introduced the function of each gene individually, enabling an in-depth understanding of their potential roles and interactions within the biological system under investigation. For the conclusion part, the report reiterated the key finding that the combination of *CCNA2*, *CCNB1/2*, *CCND2*, *CCNE1/2*, *CCNF*, *CCNG1/2*, and *CCNA* is considered to be a potential prognostic signature (Figure 6).

## 4 Discussion and Conclusion

The development of prognostic signatures in the field of medicine has always been a prime area of interest due to its potential to significantly improve patient outcomes [3, 4, 34, 35, 36, 37]. Prognostic signatures, especially those that can holistically consider multi-gene factors, have the potential to greatly influence decision-making processes in medical treatments and interventions. In cancer, the ability to stratify patients based on potential responses to treatment or the likelihood of disease recurrence can significantly enhance patients' care and improve outcomes [37, 38]. In our study, we propose MulMarker, a GPT-assisted comprehensive framework to identify and evaluate potential multi-gene prognostic signatures across various diseases. To the best of our knowledge, MulMarker is the first tool to identify potential multi-gene prognostic signatures across diseases and datasets. MulMarker allows for screening candidate genes, building risk models, and validating the performance of the identified prognostic signatures. Through extensive experiments on the METERBRIC breast cancer datasets, MulMarker identified a cell-cycle-related prognostic signature for breast cancer, which is validated by the results of Liu et al. [22]. Besides, MulMarker integrates GPT-3.5-Turbo to develop an interactive chatbot to address user queries about inputs, algorithms, and analysis details. This integration not only aids in solving technical doubts but also makes the tool more feasible and user-friendly. More-

## A

### Analysis Results

| Screen   | Formula   | Train Group | Test Group | Total Dataset  |                |                     |                     |        |           |              |           |
|--|-----------|-------------|------------|----------------|----------------|---------------------|---------------------|--------|-----------|--------------|-----------|
| The result of univariate Cox regression analysis is: |           |             |            |                |                |                     |                     |        |           |              |           |
| covariate  | coef      | exp(coef)   | se(coef)   | coef lower 95% | coef upper 95% | exp(coef) lower 95% | exp(coef) upper 95% | exp to | z         | p            | log2(p)   |
| CCNA1  | -0.147645 | 0.862738    | 0.110688   | -0.364588      | 0.069299       | 0.694482            | 1.071757            | 0.0    | -1.333884 | 1.822419e-01 | 2.456073  |
| CCNA2  | 0.221000  | 1.247324    | 0.051506   | 0.120051       | 0.321950       | 1.127554            | 1.379816            | 0.0    | 4.290782  | 1.780451e-05 | 15.777397 |
| CCNB1  | 0.253325  | 1.288301    | 0.061533   | 0.132723       | 0.373927       | 1.141933            | 1.453431            | 0.0    | 4.116902  | 3.839995e-05 | 14.668536 |
| CCNB2  | 0.174127  | 1.190207    | 0.038298   | 0.099064       | 0.249190       | 1.104137            | 1.282986            | 0.0    | 4.546628  | 5.451221e-06 | 17.484989 |
| CCNB3  | -0.001649 | 0.998353    | 0.129429   | -0.255324      | 0.252027       | 0.774665            | 1.286631            | 0.0    | -0.012737 | 9.898376e-01 | 0.014736  |
| CCNC   | -0.032722 | 0.967807    | 0.046902   | -0.124649      | 0.059205       | 0.882806            | 1.066993            | 0.0    | -0.097664 | 4.853874e-01 | 1.042792  |
| CCND1  | 0.034461  | 1.035061    | 0.034737   | -0.033623      | 0.102544       | 0.966936            | 1.107986            | 0.0    | 0.992042  | 3.211772e-01 | 1.638558  |
| CCND2  | -0.293507 | 0.745644    | 0.051213   | -0.393882      | -0.193131      | 0.674433            | 0.824374            | 0.0    | -5.731100 | 9.978159e-09 | 26.578579 |
| CCND3  | -0.013200 | 0.986878    | 0.088782   | -0.187218      | 0.160801       | 0.829263            | 1.174451            | 0.0    | -0.148776 | 8.817307e-01 | 0.181590  |
| CCNE1  | 0.124334  | 1.132394    | 0.039895   | 0.046141       | 0.202526       | 1.047222            | 1.224492            | 0.0    | 3.116533  | 1.829912e-03 | 9.094010  |
| CCNE2  | 0.231458  | 1.260437    | 0.049789   | 0.133874       | 0.329043       | 1.143249            | 1.389637            | 0.0    | 4.648790  | 3.338878e-06 | 18.192205 |
| CCNF   | 0.317103  | 1.373144    | 0.060204   | 0.199105       | 0.451501       | 1.220310            | 1.545119            | 0.0    | 5.267121  | 1.385806e-07 | 22.782778 |
| CCNG1  | -0.225664 | 0.797986    | 0.057608   | -0.338574      | -0.112755      | 0.712786            | 0.893369            | 0.0    | -3.917251 | 8.956459e-05 | 13.446712 |
| CCNG2  | -0.187617 | 0.828932    | 0.054210   | -0.293866      | -0.081368      | 0.745376            | 0.921854            | 0.0    | -3.460957 | 5.382592e-04 | 10.859411 |
| CCNH   | -0.175158 | 0.839325    | 0.065944   | -0.304404      | -0.045911      | 0.737563            | 0.955127            | 0.0    | -2.656176 | 7.903236e-03 | 6.983341  |

Show Less

Download

## B

### Analysis Results

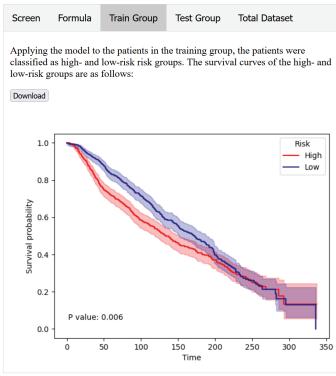
| Screen   | Formula | Train Group | Test Group | Total Dataset |
|--|---------|-------------|------------|---------------|
| A predictive model is constructed using the selected genes and their respective hazard ratios. The selection of these genes is based on a univariate Cox regression analysis, where only those with a p-value of less than 0.05 are chosen. The hazard ratios are determined by exponentiating the coefficients (exp(coef)) of the corresponding genes from the results of the multivariate Cox regression analysis. The formula is: |         |             |            |               |
| 1.124*CCNA2 + 1.160*CCNB1 + 0.979*CCNB2 + 0.777*CCND2 + 0.900*CCNE1 + 1.056*CCNE2 + 0.928*CCNG1 + 0.896*CCNG2 + 0.865*CCNH   |         |             |            |               |

Show Less

Download

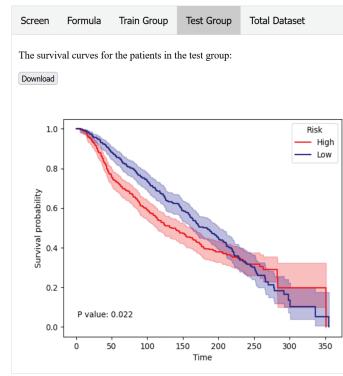
## C

### Analysis Results



## D

### Analysis Results



## E

### Analysis Results

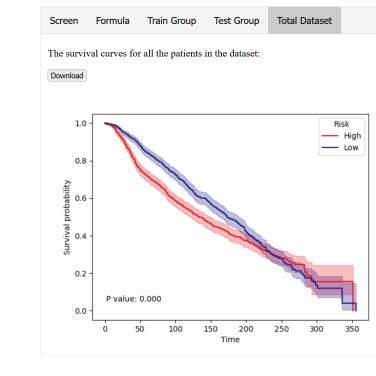


Figure 5: Analysis results. (A) Results for the univariate Cox regression analysis. (B) Results for the multivariate Cox regression analysis. (C) Survival curve for the high- and low-risk groups in the training group. (D) Survival curve for the high and low-risk groups in the test group. (E) Survival curve for the high- and low-risk groups in the total data.

## Explanations

MulMarker is a comprehensive framework for identifying potential multi-gene prognostic signatures across various diseases. With quantified data and clinical information of patients, MulMarker will screen candidate genes as a potential prognostic signature and use the KM survival analysis and log-rank test to evaluate its performance.

In the first step of MulMarker, the patients are randomly divided into a training group and a test group. The training group is used to identify candidate genes through univariate Cox regression analysis. The candidate genes identified in this process are: CCNA2, CCNB1, CCNB2, CCND2, CCNE1, CCNE2, CCNF, CCNG1, CCNG2, and CCNH.

After the candidate genes are identified, a multivariate Cox regression model is constructed using these genes. The risk model formula is:  $1.124 * CCNA2 + 1.160 * CCNB1 + 0.979 * CCNB2 + 0.777 * CCND2 + 0.900 * CCNE1 + 1.056 * CCNE2 + 1.065 * CCNF + 0.928 * CCNG1 + 0.896 * CCNG2 + 0.865 * CCNH$ .

Next, the risk value for each patient in the training group is calculated based on this risk model formula. The patients are then divided into high-risk and low-risk groups based on the median of the risk values. If the risk value is higher than the median, the patient is classified as high-risk. If the risk value is equal to or lower than the median, the patient is classified as low-risk. The median value in the training group is 72.054.

The same risk model formula and risk threshold are applied to the patients in the test group. The median value in the test group is 71.855.

Finally, the risk model is also applied to the total group of patients. The median value in the total group is 71.954.

To evaluate the performance of the risk model, survival analysis and log-rank test are used. The p-values of the log-rank test in the training, test, and total groups are 0.006, 0.022, and 0.0 respectively.

Based on the results, the candidate genes identified by MulMarker, including CCNA2, CCNB1, CCNB2, CCND2, CCNE1, CCNE2, CCNF, CCNG1, CCNG2, and CCNH, are a potential prognostic signature for patient stratification.

These genes play various roles in the progression of diseases. Here is a brief introduction to the role of each gene:

1. CCNA2: This gene encodes a regulatory protein involved in the progression of the cell cycle. It plays a critical role in cell division and is associated with tumor proliferation.

2. CCNB1: This gene is involved in regulating the cell cycle and is essential for the progression through the G2 phase. Abnormal expression of CCNB1 is associated with uncontrolled cell proliferation and is commonly found in various cancers.

3. CCNB2: Similar to CCNB1, CCNB2 is involved in regulating the cell cycle, particularly in the G2 phase. It is associated with cell division and is often dysregulated in cancer cells.

4. CCND2: This gene encodes a protein that regulates the cell cycle by promoting progression through the G1 phase. Aberrant expression of CCND2 is commonly observed in cancer cells and is associated with increased cell proliferation.

5. CCNE1: CCNE1 is involved in the regulation of the cell cycle, particularly in the G1/S transition. Overexpression of CCNE1 has been linked to uncontrolled cell proliferation and is associated with poor prognosis in various cancers.

6. CCNE2: Similar to CCNE1, CCNE2 is involved in regulating the G1/S transition of the cell cycle. Dysregulation of CCNE2 is commonly observed in cancer cells and is associated with increased cell proliferation and poor prognosis.

7. CCNF: This gene encodes a regulatory protein that controls the cell cycle by promoting the transition from the G2 phase to the M phase. Dysregulation of CCNF has been implicated in tumor progression and is associated with aggressive cancer phenotypes.

8. CCNG1: CCNG1 plays a role in regulating the cell cycle by inhibiting cyclin-dependent kinase activity. It acts as a tumor suppressor gene and its downregulation is associated with increased cell proliferation and poor prognosis in various cancers.

9. CCNG2: Similar to CCNG1, CCNG2 functions as a negative regulator of the cell cycle by inhibiting cyclin-dependent kinases. Its downregulation has been associated with increased cell proliferation and poor patient prognosis in cancer.

10. CCNH: This gene encodes a regulatory protein that is involved in the control of the cell cycle and DNA repair. Dysregulation of CCNH has been implicated in tumor progression and is associated with increased cell proliferation.

In conclusion, the candidate genes identified by MulMarker, including CCNA2, CCNB1, CCNB2, CCND2, CCNE1, CCNE2, CCNF, CCNG1, CCNG2, and CCNH, form a potential prognostic signature for patient stratification. These genes have been found to play important roles in regulating the cell cycle and are associated with tumor proliferation and prognosis in various cancers. Further validation and clinical studies are warranted to confirm their utility as prognostic markers and potential therapeutic targets for patient stratification.

*Note: The explanation for the result of MulMarker is generated by machine and is for reference only.*

Figure 6: The tailored report based on the analysis results.

over, MulMarker can generate a detailed report based on the analysis results, providing a more comprehensive and targeted explanation for the analysis results. This is also an attempt to automatically generate medical reports. Such detailed reports can be crucial in translational medicine where communication of complex medical data to clinicians and patients is vital.

However, as with all methodologies, certain considerations need to be acknowledged. While MulMarker is adaptable to various diseases, the accuracy and validity of its findings depend heavily on the quality and comprehensiveness of the input datasets. Besides, the computational identification of prognostic signatures is only the first step. These findings need subsequent validation in biological experiments, clinical trials, and large patient cohorts to assess their real-world efficacy. Moreover, using LLMs to address user queries and generate tailored reports has shown mixed results. It is not always stable and can produce answers that are unsuitable or off-target. However, with the rapid advancements in AI, we believe that many of these challenges will be addressed in the future, making them more reliable for such applications.

In conclusion, MulMarker provides a comprehensive framework to identify potential multi-gene prognostic signatures. By employing LLMs to address user queries and generate the corresponding report, it explores the possibility of integrating cutting-edge AI solutions in prognostic research. Still, it's essential to recognize the need for biological validation of these identified prognostic signatures to confirm their efficacy.

## References

1. Sechidis K, Papangelou K, Metcalfe PD, Svensson D, Weatherall J, Brown G. Distinguishing prognostic and predictive biomarkers: an information theoretic approach. *Bioinformatics*. 2018;34(19):3365-76.
2. Kerr DJ, Yang L. Personalising cancer medicine with prognostic markers. *EBioMedicine*. 2021;72.
3. Michiels S, Ternès N, Rotolo F. Statistical controversies in clinical research: prognostic gene signatures are not (yet) useful in clinical practice. *Annals of Oncology*. 2016;27(12):2160-7.
4. Nalejska E, Maczyńska E, Lewandowska MA. Prognostic and predictive biomarkers: tools in personalized oncology. *Molecular diagnosis & therapy*. 2014;18:273-84.
5. An Y, Wang Q, Zhang G, Sun F, Zhang L, Li H, et al. OSlihc: An online prognostic biomarker analysis tool for hepatocellular carcinoma. *Frontiers in Pharmacology*. 2020;11:875.
6. Aguirre-Gamboa R, Gomez-Rueda H, Martínez-Ledesma E, Martínez-Torteya A, Chacolla-Huaranga R, Rodriguez-Barrientos A, et al. SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. *PloS one*. 2013;8(9):e74250.
7. Lánczky A, Győrffy B. Web-based survival analysis tool tailored for medical research (KMplot): development and implementation. *Journal of medical Internet research*. 2021;23(7):e27633.
8. Győrffy B, Hatzis C, Sanft T, Hofstatter E, Aktas B, Pusztai L. Multigene prognostic tests in breast cancer: past, present, future. *Breast cancer research*. 2015;17(1):1-7.
9. An Y, Wang Q, Sun F, Zhang G, Wang F, Zhang L, et al. OSucs: An online prognostic biomarker analysis tool for uterine carcinosarcoma. *Genes*. 2020;11(9):1040.
10. Xie L, Wang Q, Yan Z, Han Y, Ma X, Li H, et al. OSgc: A Web Portal to Assess the Performance of Prognostic Biomarkers in Gastric Cancer. *Frontiers in Oncology*. 2022;12:856988.
11. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *Journal of Medical Systems*. 2023;47(1):33.
12. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nature Medicine*. 2023;1-11.
13. Radford A, Narasimhan K, Salimans T, Sutskever I, et al. Improving language understanding by generative pre-training. *OpenAI Blog*. 2018.
14. ZHAO X, LU J, DENG C, ZHENG C, WANG J, CHOWDHURY T, et al. Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey. *arXiv preprint arXiv:230518703*. 2023.
15. Guo CR, Mao Y, Jiang F, Juan CX, Zhou GP, Li N. Computational detection of a genome instability-derived lncRNA signature for predicting the clinical outcome of lung adenocarcinoma. *Cancer medicine*. 2022;11(3):864-79.
16. Guo M, Wang SM. Genome instability-derived genes are novel prognostic biomarkers for triple-negative breast

- cancer. *Frontiers in Cell and Developmental Biology*. 2021;9:701073.
- 17. Jiang H, Xu S, Chen C. A ten-gene signature-based risk assessment model predicts the prognosis of lung adenocarcinoma. *BMC cancer*. 2020;20(1):1-11.
  - 18. Liu C, Li Y, Wei M, Zhao L, Yu Y, Li G. Identification of a novel glycolysis-related gene signature that can predict the survival of patients with lung adenocarcinoma. *Cell Cycle*. 2019;18(5):568-79.
  - 19. Xie H, Xie C, et al. A six-gene signature predicts survival of adenocarcinoma type of non-small-cell lung cancer patients: a comprehensive study based on integrated analysis and weighted gene coexpression network. *BioMed Research International*. 2019;2019.
  - 20. Zhang L, Zhang Z, Yu Z. Identification of a novel glycolysis-related gene signature for predicting metastasis and survival in patients with lung adenocarcinoma. *Journal of translational medicine*. 2019;17(1):1-13.
  - 21. Zhang X, Lam TW, Ting HF. Genome Instability-Derived Genes as a Novel Prognostic Signature for Lung Adenocarcinoma. *Frontiers in Cell and Developmental Biology*;11:1224069.
  - 22. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *cell*. 2011;144(5):646-74.
  - 23. Liu NQ, Cao WH, Wang X, Chen J, Nie J. Cyclin genes as potential novel prognostic biomarkers and therapeutic targets in breast cancer. *Oncology Letters*. 2022;24(4):1-13.
  - 24. Carlson J. *Redis in action*. Simon and Schuster; 2013.
  - 25. Zhang S. Nearest neighbor selection for iteratively kNN imputation. *Journal of Systems and Software*. 2012;85(11):2541-52.
  - 26. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*. 2011;12:2825-30.
  - 27. Davidson-Pilon C. lifelines: survival analysis in Python. *Journal of Open Source Software*. 2019;4(40):1317.
  - 28. Wang J, Shi E, Yu S, Wu Z, Ma C, Dai H, et al. Prompt engineering for healthcare: Methodologies and applications. *arXiv preprint arXiv:230414670*. 2023.
  - 29. Hu Y, Ameer I, Zuo X, Peng X, Zhou Y, Li Z, et al. Zero-shot clinical entity recognition using chatgpt. *arXiv preprint arXiv:230316416*. 2023.
  - 30. Grinberg M. *Flask web development: developing web applications with python*. "O'Reilly Media, Inc."; 2018.
  - 31. Pereira B, Chin SF, Rueda OM, Vollan HKM, Provenzano E, Bardwell HA, et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nature communications*. 2016;7(1):11479.
  - 32. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery*. 2012;2(5):401-4.
  - 33. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling*. 2013;6(269):pl1-1.
  - 34. Bodaghi A, Fattahi N, Ramazani A. Biomarkers: Promising and valuable tools towards diagnosis, prognosis and treatment of Covid-19 and other diseases. *Heliyon*. 2023.
  - 35. Tousignant-Laflamme Y, Houle C, Cook C, Naye F, LeBlanc A, Décarie S. Mastering prognostic tools: an opportunity to enhance personalized care and to optimize clinical outcomes in physical therapy. *Physical Therapy*. 2022;102(5):pzac023.
  - 36. Burska A, Roget K, Blits M, Soto Gomez L, Van De Loo F, Hazelwood L, et al. Gene expression analysis in RA: towards personalized medicine. *The pharmacogenomics journal*. 2014;14(2):93-106.
  - 37. Simon R. Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. *Personalized medicine*. 2010;7(1):33-47.
  - 38. Méndez Hernández R, Ramasco Rueda F. Biomarkers as prognostic predictors and therapeutic guide in critically ill patients: Clinical Evidence. *Journal of Personalized Medicine*. 2023;13(2):333.