Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Haoling Xu Date

Association between drug usage time and gene expression-based pathway score in HNSCC patients

By

Haoling Xu

Master of Public Health

Department of Biostatistics and Bioinformatics

Zhaohui (Steve) Qin, PhD

(Thesis Advisor)

Yi-Juan Hu, PhD

(Reader)

Association between drug usage time and gene expression-based pathway score in HNSCC patients

By

Haoling Xu

B.S. in Pharmaceutical Preparation

China Pharmaceutical University (CPU), School of Pharmacy, 2021

Thesis Committee Chair: Zhaohui (Steve) Qin, PhD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Public Health

in Biostatistics

2023

Abstract

Association between drug usage time and gene expression-based pathway score in HNSCC patients

By Haoling Xu

Head and neck cancer is a complex cancer that could appear in several sites in the throat and head. However, due to the heterogeneity of cancer at the molecular level, it's not easy to treat all HNSCC patients in the same way. In this study, we aimed to apply a newly developed method iPath to calculate an individual enrichment score of HNSCC patients based on different KEGG pathways. We related the score to patients' drug usage time to find whether there is a significant difference between the high-score group and the low-score group. We used unpaired two-sample Wilcoxon tests and multiple test adjustments. Pathway KEGG_ARACHIDONIC_ACID_METABOLISM and pathway KEGG_LONG_TERM_DEPRESSION showed significant differences. The Kaplan-Meier analysis also indicated that KEGG_ARACHIDONIC_ACID_METABOLISM might be a prognostic biomarker pathway for head and neck squamous cell carcinoma patients who take the drug Carboplatin.

Association between drug usage time and gene expression-based pathway score in HNSCC patients

By

Haoling Xu

B.S. in Pharmaceutical Preparation

China Pharmaceutical University (CPU), School of Pharmacy, 2021

Thesis Committee Chair: Zhaohui (Steve) Qin, PhD

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Public Health in Biostatistics

2023

Acknowledgement

I would like to thank Emory University and the Rollins School of Public Health for my two years of study in biostatistics. I'm grateful to meet everyone here. I'm also grateful for having Steve Qin as my thesis professor, who has professional knowledge in biostatistics and bioinformatics. He teaches me a lot. Then I would like to express my gratitude to my friends at Emory University, as well as those who chat with me online, for their emotional support and guidance. Lastly, I extend my appreciation to both my parents and myself.

Table of Contents

1. Introduction

Head and neck cancer is a complex cancer that could appear in several sites in the throat and head. It is estimated to have over 600,000 cases annually around the world [1]. Among head and neck cancer, head and neck squamous cell carcinoma (HNSCC) is the most common histological type [2]. There are three main treatments of HNSCC, immunotherapy, chemotherapy, and molecular targeted therapy. 5-FU, Cisplatin, and docetaxel are the most often used drugs in clinical studies [3]. However, due to the heterogeneity of cancer at the molecular level, it's not easy to treat all HNSCC patients in the same way. Even if the main treatments are used in combination, these treatments may prove ineffective and adversely affect the patient's quality of life [4]. Therefore, there is an urgent need to evaluate how patients respond to current treatment and find if there are common characteristics to help provide better individual treatment.

One promising approach is to study the pathways that are related to tumor growth and progression. These pathways are complex networks of molecular interactions that control cell behavior. The regulation of critical functions such as growth, proliferation, and cell-death pathways can be poorly managed in cancer [5].Besides, pathways provide advantages over individual genes in aggregating molecular events, results interpretation, and finding potential mechanisms [6].

In this study, we applied a newly developed method iPath to calculate an individual enrichment score of HNSCC patients based on pathways. And we relate the pathway score to their drug usage time to find whether there is a significant difference in drug response and gene expression level.

2. Methods

2.1 Data

2.1.1 Gene expression data

The Cancer Genome Atlas (TCGA) provides genomic information and treatment information of different cancer patients. Data used in this study are downloaded from the portal of the Broad Institute GDAC Firehose analysis using TCGAbiolinks Bioconductor package (version 2.25.3) [7]. Data used in this study are RSEM (RNA-Seq by Expectation-Maximization) normalized RNAseq data.

2.1.2 Clinical data

Drug usage information are corresponded clinical data of each patient from TCGA.

2.1.3 Pathway data

KEGG pathway data in the C2 collection of MSigdb database were used [8]. This collection includes 183 manually curated KEGG pathways [9].

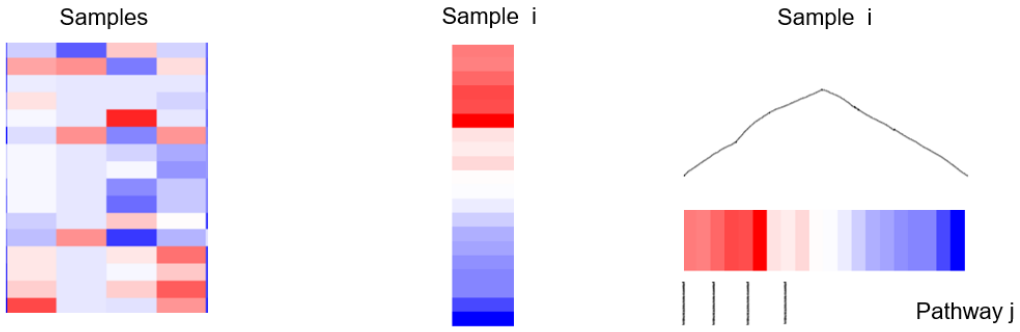
2.2 Data analysis

R software (version 4.2.0) were used in data analysis and plotting.

Gene expression data may come from one patient several times, which could be found from the TCGA patient barcode. We selected patients with tumor samples and retained their 12-character barcode for further usage. Drug usage time data were selected from the drug information dataset and were merged with gene expression data. Patients who didn't use the drug or with 0 drug usage time were deleted. Duplicated samples were also deleted. After data processing, there are 84 patients who used Cisplatin and 52 patients who used Carboplatin.

2.3 Calculation of individual enrichment score

The individual enrichment score (iES) of each HNSCC patient were calculated based on pathways by using iPath Bioconductor package (version 1.4.2) [10]. The method proposed by Su et.al (2021) aims to detect gene sets and pathways that have significant deviations from norms, which is effective in finding highly predictive biomarkers for clinical outcomes. The core of iPath is similar to the calculation of enrichment score (ES) in GSEA. Firstly, it has i samples and each sample has j genes. It denotes RNA-seq expression matrix as $Y = \{y_{ij}\}$ and calculate the mean $\bar{y}_j$ and standard deviation s_j. Then it calculates the z score z_{ij} for each gene in each sample. Genes are ranked in descending order based on the absolute value of z-score $|z_{ij}|$. The iES score is calculated by walking down the ranked list and a running-sum statistic is accumulated. Fig 1 shows how this 'iPath' method works.



(a) Normalized samples (b) Full rank gene list (c) Calculate iES score

Fig 1 calculation of individual enrichment score

2.4 Statistical analysis

The median value of the iES were used to divide patients into high score group and low score group according to iES calculated from different KEGG pathways. Unpaired two-samples Wilcoxon tests were used to the analysis of comparisons between high score and low score

groups at the significance level of 0.05. Multiple test adjustment was used to control the false discovery rate after tests. Benjamini-Hochberg procedure was applied to get q-values and these q-values were compared to a threshold of 0.05 to determine whether it's significant.

2.5 Kaplan-Meier Analysis

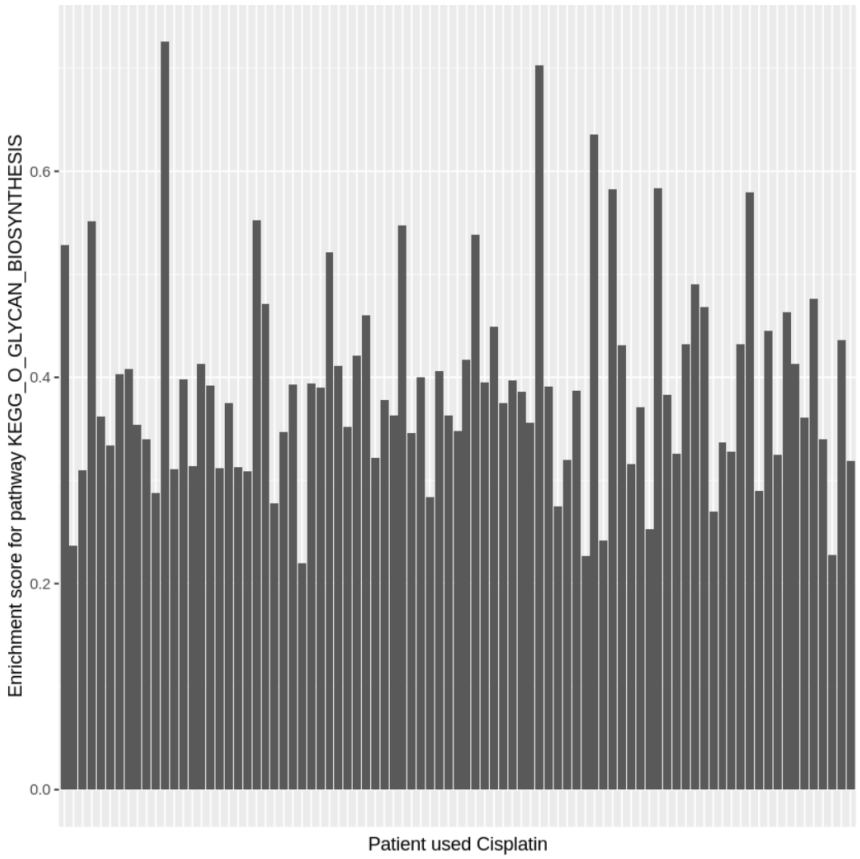
After multiple test adjustment, pathways that still shows significant difference in drug usage time between groups were selected. Kaplan-Meier analysis was applied to test if there were significance difference in survival days between high score group and low score group, with a log-rank p-value less than 0.05. Overall survival days (OS) were used in this Kaplan-Meier analysis.

3. Results

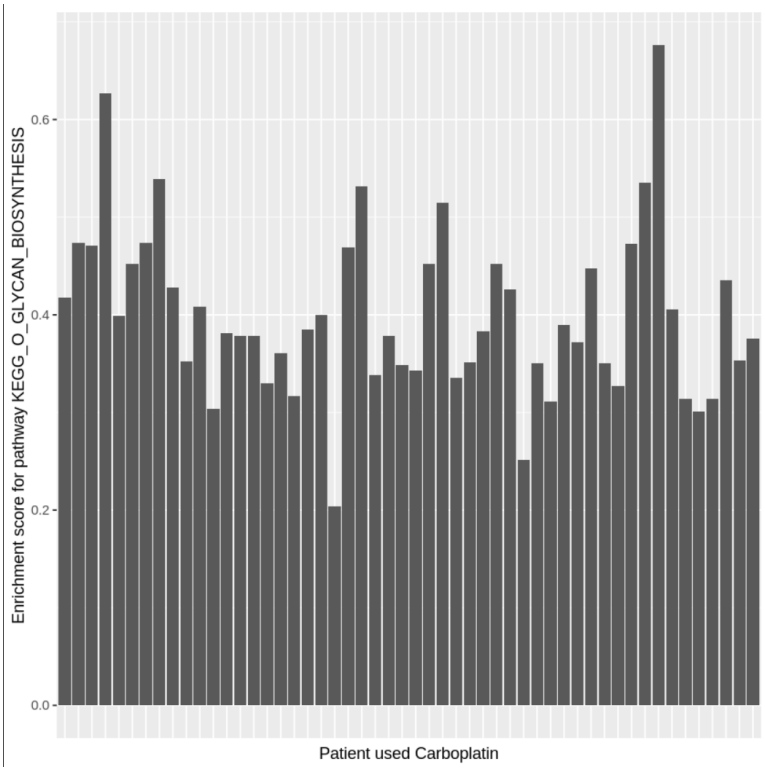
3.1 Visualization of individual enrichment score (iES) for KEGG pathways

3.1.1 Bar plot of individual enrichment score (iES) for KEGG pathways among patients

The bar plots described the levels of individual enrichment scores for KEGG pathways among HNSCC patients. Pathway KEGG_O_GLYCAN_BIOSYNTHESIS was selected as an example to show the distribution of the iES score. Figure 2a showed the bar plot of the iES of patients who used Cisplatin and Figure 2b was the bar plot of the iES of patients who used Carboplatin. The bar plots showed the same pathway may have different patterns in gene expression levels in each patient. A higher score means the gene expression level of this pathway could deviate more from the norm among samples that were included, as a z score was used to rank genes in the calculation of iES.



(a) Bar plot of iES scores for pathway KEGG_O_GLYCAN_BIOSYNTHESIS of patients used Cisplatin



(b) Bar plot of iES scores for pathway KEGG_O_GLYCAN_BIOSYNTHESIS of patients used Carboplatin

Figure 2 Visualization of individual enrichment score (iES) for one KEGG pathway among HNSCC patients

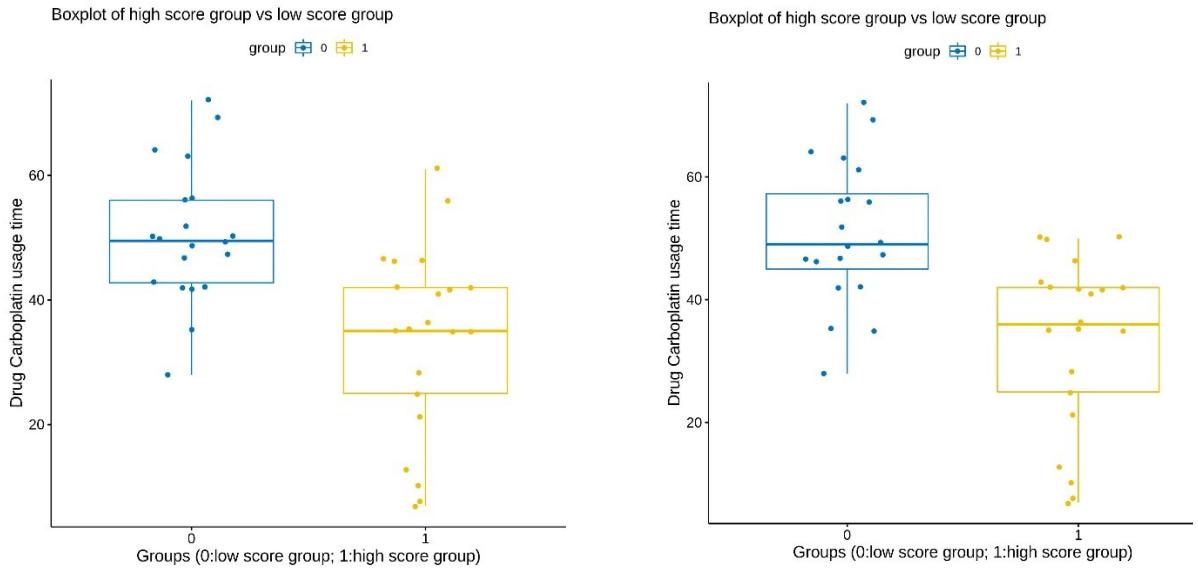
3.2 Comparison in drug usage time between high score group and low score group

3.2.1 Patients who used Cisplatin

After multiple test adjustment, there is no significant pathway among patients who took drug Cisplatin.

3.2.2 Patients who used Carboplatin

After multiple test adjustment, two pathways showed significant difference in drug usage time, which is pathway KEGG_ARACHIDONIC_ACID_METABOLISM and pathway KEGG_LONG_TERM_DEPRESSION . Figure 3a and Figure 3b were the boxplot of drug usage time between high score group and low score group. These pathways may be associated with the treatment of using Carboplatin.



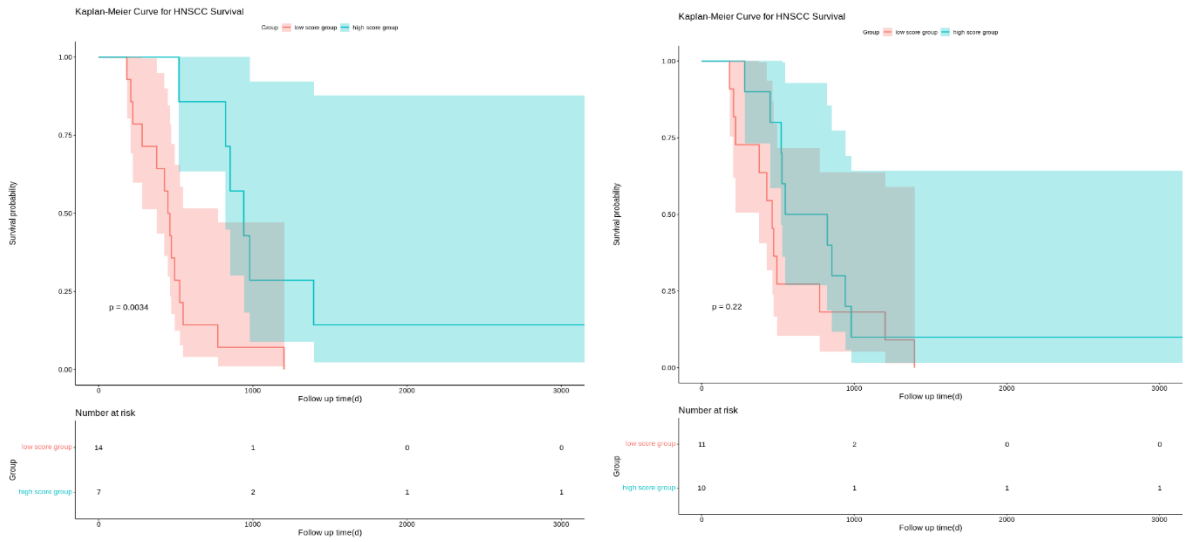
(a)KEGG_ARACHIDONIC_ACID_METABOLISM (b) KEGG_LONG_TERM_DEPRESSION

Fig 3 Pathways showing significant difference in drug Carboplatin usage time

3.3 Kaplan-Meier analysis of significant pathways

We tested the above significant pathways by using Kaplan-Meier analysis.

KEGG_ARACHIDONIC_ACID_METABOLISM showed a significant difference in survival probability between 2 groups (p=0.034), while KEGG_LONG_TERM_DEPRESSION showed no significant difference.



(a) (b)

KEGG_ARACHIDONIC_ACID_METABOLISM KEGG_LONG_TERM_DEPRESSION

Fig 4 Kaplan-Meier Analysis of Significant Pathways

4. Discussion

From the above study, we found there are two pathways that showed significant difference in drug usage time for patients who used drug Cisplatin. And pathway KEGG_ARACHIDONIC_ACID_METABOLISM showed a significant difference in survival probability between 2 groups. Pathway KEGG_ARACHIDONIC_ACID_METABOLISM might be a prognostic biomarker pathway for head and neck squamous cell carcinoma patients who take drug Carboplatin. The biological explanation for this finding is that the arachidonic acid metabolism pathway plays a crucial role in cardiovascular biology and other diseases. The enzymes produced by this pathway may also have a regulatory effect on the development of tumors, such as the arachidonic acid-catalyzed lipoxygenase[11].

Our study may have some limitations. One potential limitation of using clinical data is that errors may have occurred during the recording process, which could impact the accuracy and precision of the study results. In addition, specific pathways associated with the pathogenesis of HNSCC were not taken into account.

Reference

[1] Shanthi Marur, Arlene A. Forastiere, Head and Neck Cancer: Changing Epidemiology, Diagnosis, and Treatment, Mayo Clinic Proceedings,Volume 83, Issue 4,2008,Pages 489-501, ISSN 0025-6196, https://doi.org/10.4065/83.4.489. (https://www.sciencedirect.com/science/article/pii/S0025619611607064)

[2] Monica Ramos, Sergi Benavente & Jordi Giralt (2010) Management of squamous cell carcinoma of the head and neck: updated European treatment recommendations, Expert Review of Anticancer Therapy, 10:3, 339-344, DOI: 10.1586/era.10.6

[3] Goel, B., Tiwari, A. K., Pandey, R. K., Singh, A. P., Kumar, S., Sinha, A., Jain, S. K., & Khattri, A. (2022). Therapeutic approaches for the treatment of head and neck squamous cell carcinoma-An update on clinical trials. Translational oncology, 21, 101426. https://doi.org/10.1016/j.tranon.2022.101426

[4] von Witzleben A, Wang C, Laban S, Savelyeva N, Ottensmeier CH. HNSCC: Tumour Antigens and Their Targeting by Immunotherapy. Cells. 2020;9(9):2103. Published 2020 Sep 15. doi:10.3390/cells9092103

[5] Colaprico A, Olsen C, Bailey MH, et al. Interpreting pathways to discover cancer driver genes with Moonlight. Nat Commun. 2020;11(1):69. Published 2020 Jan 3. doi:10.1038/s41467-019-13803-0

[6] Creixell P, Reimand J, Haider S, et al. Pathway and network analysis of cancer genomes. Nat Methods. 2015;12(7):615-621. doi:10.1038/nmeth.3440

[7] Antonio Colaprico, Tiago C. Silva, Catharina Olsen, Luciano Garofano, Claudia Cava,

Davide Garolini, Thais S. Sabedot, Tathiane M. Malta, Stefano M. Pagnotta, Isabella Castiglioni, Michele Ceccarelli, Gianluca Bontempi, Houtan Noushmehr, TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data, Nucleic Acids Research, Volume 44, Issue 8, 5 May 2016, Page e71, https://doi.org/10.1093/nar/gkv1507

[8] Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. Cell systems, 1(6), 417-425.

[9] Minoru Kanehisa, Susumu Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes, Nucleic Acids Research, Volume 28, Issue 1, 1 January 2000, Pages 27–30, https://doi.org/10.1093/nar/28.1.27

[10] Kenong Su, Qi Yu, Ronglai Shen, Shi-Yong Sun, Carlos S. Moreno, Xiaoxian Li, Zhaohui S. Qin,Pan-cancer analysis of pathway-based gene expression pattern at the individual level reveals biomarkers of clinical prognosis,Cell Reports Methods,Volume 1, Issue 4,2021,100050,ISSN 2667-2375,https://doi.org/10.1016/j.crmeth.2021.100050

[11] Pidgeon GP, Lysaght J, Krishnamoorthy S, et al. Lipoxygenase metabolism: roles in tumor progression and survival. Cancer Metastasis Rev. 2007;26(3-4):503-524. doi:10.1007/s10555-007-9098-3