FINAL SCRIPT: APPROVED FOR FILMING

Submission ID #: 67941
Scriptwriter Name: Poornima G
Project Page Link: https://review.jove.com/files_upload.php?src=20722053

Title: Comparison of Predictive Performance of Three Lymph Node Staging Systems in Colorectal Signet Ring Cell Carcinoma Based on Machine Learning Model

Authors and Affiliations:
Jinyan Jia[1,2*], Zixuan Yu[1,2,3*], Maorun Zhang[1, 2], Fang Hu[3], Gang Liu[1,2]

[1]Department of General Surgery, Tianjin Medical University General Hospital.
[2]Tianjin Key Laboratory of Precise Vascular Reconstruction and Organ Function Repair.
[3]Department of Nursing, Tianjin Medical University General Hospital.

[*]These authors contributed equally

Corresponding Authors:

Gang Liu	lg1059@tmu.edu.cn
Fang Hu	1430190119@qq.com

Email Addresses for All Authors:

Jinyan Jia	yzzsjc202@163.com
Zixuan Yu	2423256156@qq.com
Maorun Zhang	zzuzmr@163.com
Gang Liu	lg1059@tmu.edu.cn
Fang Hu	1430190119@qq.com

Author Questionnaire

1. Microscopy: Does your protocol require the use of a dissecting or stereomicroscope for performing a complex dissection, microinjection technique, or something similar? **NO**

2. Software: Does the part of your protocol being filmed include step-by-step descriptions of software usage? **Yes, all done**

3. Filming location: Will the filming need to take place in multiple locations? **No**

Current Protocol Length

Number of Steps:	16
Number of Shots:	42 (41 SC)

Introduction

Videographer: Obtain headshots for all authors available at the filming location.

1.1. **Gang Liu:** Our research evaluates three lymph node staging systems in colorectal signet ring cell carcinoma using machine learning and competing risk models to optimize prognostic accuracy and survival prediction.

1.1.1. INTERVIEW: Named talent says the statement above in an interview-style shot, looking slightly off-camera. *Suggested B-roll: 2.3.1*

What technologies are currently used to advance research in your field?

1.2. **Jinyan Jia:** Bioinformatics methods, including machine learning, competing risk models and Kaplan-Meier survival estimation, are used to enhance survival prediction and lymph node classification accuracy.

1.2.1. INTERVIEW: Named talent says the statement above in an interview-style shot, looking slightly off-camera. *Suggested B-roll: 4.1.1*

What research questions will your laboratory focus on in the future?

1.3. **Zixuan Yu :** Extending follow-up periods, validating in diverse populations, refining prognostic nomograms, and exploring molecular traits of colorectal signet ring cell carcinoma to enhance clinical decision-making tools.

1.3.1. INTERVIEW: Named talent says the statement above in an interview-style shot, looking slightly off-camera. *Suggested B-roll: 4.2.1*

Videographer: Obtain headshots for all authors available at the filming location.

Protocol

2. Data Acquisition for Modelling

Demonstrator: Jinyan Jia

2.1. To begin, download and install SEER *(S-E-E-R)* **[1],** then obtain the statistics 8.4.3 software from the SEER database website **[2]**. Log into the software and click on **Case List Session** followed by **Data [3]** and select the **Incidence SEER Research Plus Data, 17 Registries, Nov 2022 Sub (2000-2020)** *(two thousand to twenty twenty)* database **[4]**.

2.1.1. WIDE: Talent taking a seat at the computer table.

2.1.2. SCREEN: 2.1.2 00:05-00:18.

2.1.3. SCREEN: 2.1.3.

2.1.4. SCREEN: 2.1.4.

2.2. Now, click on **Selection** followed by **Edit** and choose **Race, Sex, Year of diagnosis** equal to '2004' through '2015' **[1]**. Then, select **Site recode ICD-O-3 WHO 2008** *(site-recode-I-C-D-Oh-3-W-H-Oh-two thousand and eight)* **[2]**.

2.2.1. SCREEN: 2.2.1.

2.2.2. SCREEN: 2.2.2.

2.3. Click on **Table**, and in the available variables interface, select all the diagnosis details required **[1]**.

2.3.1. SCREEN: 2.3.1

2.4. Then, click on **Output**, name the data, **[1]** and click on **Execute** to output and save the data **[2]**.

2.4.1. SCREEN: 2.4.1.

2.4.2. SCREEN: 2.4.2.

2.5. Next, open the X-tile software, click on **File** and choose **Open [1].** Select the data file to import it into the software **[2]**. Once the data is loaded, map the variable **Censor** corresponding to survival status, the **Survival time** and **Marker1** as the variable to be analyzed, ensuring the data matches correctly **[3]**.

2.5.1. SCREEN: 2.5.1. *Video editor: If possible blur the background and keep only the X-TILE software interface in focus*

2.5.2. SCREEN: 2.5.2.

2.5.3. SCREEN: 2.5.3.

2.6. Now, click on **Do** followed by **Kaplan-Meier** and **Marker1** to perform the Kaplan-Meier survival analysis **[1]** and generate the survival curve **[2]**.

2.6.1. SCREEN: .2.6.1
2.6.2. SCREEN: 2.6.2. 00:03-00:06

NOTE: Please display step 3.1 here before 2.7

2.7. Then, randomly assign a total of 2,409 eligible patient data with SRCC to a training cohort number 1,686 and a validation cohort number 723 in a 7 to 3 ratio **[1]**. Use the provided code for random splitting **[2]**.
2.7.1. SCREEN: 2.7.1..
2.7.2. SCREEN: 2.7.2.

3. **Machine Learning Models Development and Verification**

Download and install the required versions of R-Studio and R software **[1]**. Click on **New File** and select **R Script** to create a new R programming interface **[2]**. Then, enter the relevant code in the code editor and click on **Run** to execute the code **[3]**. **NOTE**: Please move step 3.1 before 2.7
3.1.1. SCREEN: 3.1.1. **TXT: RStudio version: 2024.04.2+764; R software version: 4.4.1**
3.1.2. SCREEN: 3.1.2.
3.1.3. SCREEN: 3.1.3.

3.2. Use the provided code to screen the variables included in the machine learning models by Cox regression analysis **[1]**. Additionally, explore the impact of LODDS *(L-O-D-D-S)*, LNR *(L-N-R)*, and pN *(P-N)* staging on cancer-specific survival in SRCC patients **[2-TXT]**.
3.2.1. SCREEN: 3.2.1.
3.2.2. SCREEN: 3.2.2.
TXT: The traindata.csv is data obtained from the SEER database

3.3. Use the code to compare the prognostic prediction abilities of three lymph node systems LODDS, LNR, and pN staging **[1]** across the training, validation, and external validation cohorts **[2]**.
3.3.1. SCREEN: 3.3.1.
3.3.2. SCREEN: 3.3.2. 00:10-00:16

3.4. Then use the code to build an XGBoost *(X-G-Boost)* model **[1]** and generate bar graphs representing the relative importance of variables **[2]**. Generate receiver operating characteristic curves **[3]** and calibration curves to assess the performance of the three lymph node systems **[4]**.
3.4.1. SCREEN: 3.4.1. 00:07-00:14
3.4.2. SCREEN: 3.4.2.
3.4.3. SCREEN: 3.4.3.
3.4.4. SCREEN: 3.4.4.

3.5. Next, employ the code to build a random forest model and generate bar graphs of the relative importance of variables **[1]**. Similarly, generate receiver operating characteristic curves **[2]** and calibration curves to evaluate and compare the three lymph node systems **[3]**.
3.5.1. SCREEN: 3.5.1.
3.5.2. SCREEN: 3.5.2.
3.5.3. SCREEN: 3.5.3.

3.6. With the appropriate code, build a neural network model **[1]** and produce bar graphs of the relative importance of variables **[2]**. Generate receiver operating characteristic and calibration curves to compare the predictive performance of the three lymph node systems **[3]**.
3.6.1. SCREEN: 3.6.1.
3.6.2. SCREEN: 3.6.2.
3.6.3. SCREEN: 3.6.3.

3.7. Then, perform univariate analysis **[1]** and plot the cumulative incidence function curve using the data.csv file **[2]**. Replace "Site" with other factors to perform univariate analysis for each factor **[3]**.
3.7.1. SCREEN: 3.7.1. 00:09-00:12
3.7.2. SCREEN: 3.7.2.
3.7.3. SCREEN: 3.7.3. 00:04-00:12

3.8. For multivariate analysis, apply the code **[1]** and visualize with data1.csv *(data 1 C-S-V)* **[2]**.
3.8.1. SCREEN: 3.8.1. 00:09-00:13
3.8.2. SCREEN: 3.8.2. **TXT: Click on Export, then click Save as PDF**

3.9. Finally, plot the nomogram, receiver operating characteristic curve, and calibration curve **[1]**. Train the model using data from the training cohort and use validation and external validation cohort data to validate the model **[2]**.
3.9.1. SCREEN: 3.9.1. 00:06-00:19
3.9.2. SCREEN: 3.9.2.

Results

4. Representative Results

4.1. Based on multivariate Cox regression analysis, LNR, LODDS, and pN staging were all significantly associated with cancer-specific survival in SRCC patients **[1]**.

4.1.1. LAB MEDIA: Figure 2. *Video editor: Highlight A, B, and C sequentially*.

4.2. LNR showed the highest importance in the RF and XGBoost models **[1]**, while LODDS had the greatest predictive ability in the NN model, suggesting LODDS as the most reliable LN system overall **[2]**.

4.2.1. LAB MEDIA: Figure 3. *Video editor: Focus on the panels A, B and highlight LNR bar*

4.2.2. LAB MEDIA: Figure 3. *Video editor: Focus on the panels C and highlight "LODDS" bar in C*.

4.3. The XGBoost, RF, and NN models achieved high predictive accuracy with AUC values ranging from 0.777 to 0.851 **[1]** and calibration curves that aligned closely with the 45-degree line, confirming model reliability **[2]**.

4.3.1. LAB MEDIA: Figure 4A–C.

4.3.2. LAB MEDIA: Figure 4A–F

4.4. Competing risk model analysis identified T staging, N staging, M staging, LODDS classification, and primary tumor location as independent prognostic factors **[1]**.

4.4.1. LAB MEDIA: Figure 5. *Video editor: Sequentially Highlight the graphs*.

4.5. The competing risk nomogram demonstrated accurate 1-, 3-, and 5-year cancer-specific survival predictions **[1]**, supported by well-aligned calibration and ROC curves with AUCs above 0.75 **[2]**.

4.5.1. LAB MEDIA: Figure 6A *Video editor: Focus on the lines starting with "1-year, 3-year and 5-year"*

4.5.2. LAB MEDIA: Figure 6E, F, G

Pronunciation guide

1. SEER

Pronunciation link:
https://www.merriam-webster.com/dictionary/seer
IPA: /ˈsɪr/
Phonetic Spelling: seer

---

2. Incidence

Pronunciation link:
https://www.merriam-webster.com/dictionary/incidence
IPA: /ˈɪnsɪdəns/
Phonetic Spelling: in-suh-dns

---

3. Registry

Pronunciation link:
https://www.merriam-webster.com/dictionary/registry
IPA: /ˈrɛdʒɪstri/
Phonetic Spelling: reh-juh-stree

---

4. ICD-O-3

Pronunciation link:
No confirmed link found (ICD is commonly read as initials: I-C-D; "O" as "oh"; 3 as "three")
IPA: /ˌaɪ.si.diː.oʊ.ˈθriː/
Phonetic Spelling: eye-cee-dee-oh-three

---

5. WHO (as World Health Organization abbreviation)

Pronunciation link:
https://www.merriam-webster.com/dictionary/WHO
IPA: /ˈduː/ or /ˈhuː/ (commonly pronounced as initials “W-H-O” or said as “who”)
Phonetic Spelling: double-yoo-aych-oh

---

6. X-tile

Pronunciation link:
No confirmed link found
IPA: /ˈɛks.taɪl/
Phonetic Spelling: eks-tile

---

7. Censor

Pronunciation link:
https://www.merriam-webster.com/dictionary/censor
IPA: /ˈsɛnsɚ/
Phonetic Spelling: sen-ser

---

8. Kaplan-Meier

Pronunciation link:
https://www.howtopronounce.com/kaplan-meier
IPA: /ˈkæplən ˈmaɪɚ/
Phonetic Spelling: kap-luhn my-er

---

9. SRCC (Signet Ring Cell Carcinoma - abbreviated)

Pronunciation link:
No confirmed link found (pronounced as initials: S-R-C-C)
IPA: /ˌɛs.ɑːr.siː.siː/
Phonetic Spelling: ess-ar-see-see

---

10. R-Studio

Pronunciation link:
No confirmed link found (commonly spoken as “R Studio”)
IPA: /ɑr ˈstuːdioʊ/
Phonetic Spelling: ar stoo-dee-oh

---

11. Cox regression

Pronunciation link (for 'Cox'):
https://www.merriam-webster.com/dictionary/Cox
IPA: /kɑks/
Phonetic Spelling: koks

Pronunciation link (for 'regression'):
https://www.merriam-webster.com/dictionary/regression
IPA: /rɪˈɡrɛʃən/
Phonetic Spelling: ri-gre-shun

---

12. LODDS

Pronunciation link:
No confirmed link found (pronounced as initials: L-O-D-D-S)
IPA: /ˌɛl.oʊ.diː.diː.ɛs/
Phonetic Spelling: el-oh-dee-dee-ess

---

13. LNR

Pronunciation link:
No confirmed link found (pronounced as initials: L-N-R)
IPA: /ˌɛl.ɛn.ɑr/
Phonetic Spelling: el-en-ar

---

14. pN staging

Pronunciation link:
No confirmed link found (pronounced as initials: P-N)
IPA: /ˌpiːˈɛn/
Phonetic Spelling: pee-en

---

15. XGBoost

Pronunciation link:
https://www.howtopronounce.com/xgboost

IPA: /ˈɛks.dʒiː.buːst/
Phonetic Spelling: eks-jee-boost

16. Receiver Operating Characteristic

Pronunciation link (for 'receiver'):
https://www.merriam-webster.com/dictionary/receiver
IPA: /rɪˈsiːvɚ/
Phonetic Spelling: rih-see-ver

Pronunciation link (for 'characteristic'):
https://www.merriam-webster.com/dictionary/characteristic
IPA: /ˌkærəktəˈrɪstɪk/
Phonetic Spelling: keh-ruhk-tuh-ris-tik

17. Neural Network

Pronunciation link (for 'neural'):
https://www.merriam-webster.com/dictionary/neural
IPA: /ˈnʊrəl/ or /ˈnjʊrəl/
Phonetic Spelling: nyoo-ruhl or noo-ruhl

18. Nomogram

Pronunciation link:
https://www.merriam-webster.com/dictionary/nomogram
IPA: /ˈnɑːməˌɡræm/
Phonetic Spelling: noh-muh-gram