УДК 616.33-006.6-089-085.277.3

О. В. Серая[1], С. И. Киркилевский[2], В. Г. Дубинина[3], О. В. Лукьянчук[4], А. И. Ткаченко[3], А. А. Машуков[2,3,4], В. В. Бошкова[3], А. А. Биленко[3], А. Н. Згура[4], В. Е. Максимовский[5]

ИСПОЛЬЗОВАНИЕ РЕГРЕССИОННОГО АНАЛИЗА ДЛЯ ПРОГНОЗИРОВАНИЯ ВЫЖИВАЕМОСТИ БОЛЬНЫХ РАКОМ ЖЕЛУДКА

[1]Кафедра распределенных информационных систем и облачных технологий НТУ Харьковский национальный технический университет «Харьковский политехнический институт»;
[2]Научно-исследовательское отделение опухолей органов грудной полости Национального института рака, г. Киев;
[3]Одесский Национальный медицинский университет;
[4]Отделение абдоминальной онкологии КУ «Одесский областной онкологический диспансер»;
[5]Центр реконструктивной и пластической хирургии Одесского национального медицинского университета

Summary. Seraya O. V., Kirkilevsky S. I., Dubinina V. G., Lukyanchuk O. V., Tkachenko A. I., Mashukov A. A., Boshkova V. V., Bilenko A. A., Zgura A. N., Maksimovsky V. E.- **IMPLEMENTATION OF REGRESSION ANALYSIS TO PREDICT THE ACCURATE SURVIVAL OF PATIENTS WITH GASTRIC CANCER. -** *The Department of Distributed Information Systems and Cloud Technologies of NTU Kharkiv National Technical University "Kharkov Polytechnic Institute", Kharkov; Research department of tumors of the chest cavity of the National Cancer Institute, Kiev; Odessa National Medical University; Odessa Regional Centre for Oncology Center; 5. The Center of Reconstructive and Plastic Surgery of Odessa National Medical University. – e-mail: nymba.od@gmail.com.* The objectives - to find the prediction techniques of gastric cancer patients survival. 221 patients having been radically operated on in the abdominal department of the Odessa Oncological Center during 2007 – 2011 were examined. Their life expectancy was measured in months. Such factors as the patient's age, stomach department affected, G-differentiation of the tumor, depth of gastric wall invasion, the expression of the molecular growth factor VEGFR, expression of p53 protein, expression of her2/ new peptide, invasion into micronerves, microvascular invasion, invasion into neighboring organs, volume of lymph nodes dissection, amount of affected lymph nodes, number of organs removed / resected during gastrectomy / subtotal resection and genetic type of stomach cancer had been analyzed. By regressive analysis it has been proved that factors 1,10,13, i.e. a patient's age, tumor's invasion into neibouring organs and MBP influence upon the survival rate of the patients under study. On the basis of the data analyzed formula for formal evaluation of gastric cancer patients life expactancy was obtained. The results are preliminary.

Key words: stomach cancer, regression analysis, survival.

Реферат. Серая О. В., Киркилевский С. И., Дубинина В. Г., Лукьянчук О. В., Ткаченко А. И., Машуков А. А., Бошкова В. В., Биленко А. А., Згура А. Н., Максимовский В. Е. **ИСПОЛЬЗОВАНИЕ РЕГРЕССИОННОГО АНАЛИЗА ДЛЯ ПРОГНОЗИРОВАНИЯ ВЫЖИВАЕМОСТИ БОЛЬНЫХ РАКОМ ЖЕЛУДКА.** Цель работы – на основе применения математических методов обосновать возможность выживаемости больных раком желудка. В исследование включили 221 больного, радикально прооперированого в абдоминальном отделении КУ «Одесский областной онкологический диспансер» с 2007 по 2011 г. г.

Продолжительность жизни данной группы больных измерена в месяцах. Из проанализированных факторов, клинически влияющих на выживаемость больных раком желудка (F_1 - возраст пациента; F_2 - отдел желудка; F_3 - G-дифференцировка опухоли.; F_4 - инвазия в стенку желудка; F_5 - экспрессия молекулярного фактора роста эндотелия сосудов VEGFR; F_6 - экспрессия белка p53; F_7 - экспрессия белка her2\new; F_8 - инвазия в микронервы; F_9 - микрососудистая инвазия; F_{10} - инвазия в соседние органы; F_{11} - объем лимфодиссекции; F_{12} - сумма пораженных лимфатических узлов; F_{13} - количество органов, удаленных/резецированных во время гастрэктомии/субтотальной резекции; F_{14} - генетический тип рака желудка), строго математически повлияли факторы 1, 10, 13. Получена формула для формальной оценки продолжительности жизни больных. Результаты носят предварительный характер.

Ключевые слова: рак желудка, регрессионный анализ, выживаемость.

Реферат. Сіра О. В., Кіркілевскій С. І., Дубініна В. Г., Лук'янчук О. В., Ткаченко А. І., Машуков А. А., Бошкова В. В., Біленко А. А., Згура А. Н., Максимовський В. Е. **ВИКОРИСТАННЯ РЕГРЕСІЙНОГО АНАЛІЗУ ДЛЯ ПРОГНОЗУВАННЯ ВИЖИВАННІСТІ ХВОРИХ НА РАК ШЛУНКА.** Метою даної роботи з'явився пошук шляхів прогнозування виживання хворих на рак шлунка. У дослідження було включено 221 хворий, які були радикально прооперовані в абдомінальному відділенні КУ «Одеський обласний онкологічний диспансер» з 2007 по 2011 роки. Тривалість життя даної групи хворих була виміряна в місцях. З наведених у статті чинників, клінічно впливають на виживаність хворих на рак шлунка (вік пацієнта; відділ шлунка; G-диференціювання пухлини; інвазія в стінку шлунка; експресія молекулярного фактора росту ендотелію судин VEGFR; експресія білка p53; експресія білка her2/new; інвазія в мікронерви; мікросудинна інвазія; інвазія в сусідні органи; обсяг лімфодіссекції; сума уражених лімфатичних вузлів; кількість органів, видалених / резектованих під час резекції шлунка/субтотальної резекції; генетич кий тип раку шлунка), строго математично вплинули лише 1,10,13 чинники. Отримана формула для формальної оцінки тривалості хворих. Результати носять попередній характер.

Ключові слова: рак шлунка, регресійний аналіз, виживаність.

Регрессионный анализ – мощный и эффективный статистический метод построения математических моделей, описывающих зависимость между показателем функционирования анализируемой системы y и обусловливающими его объясняющими независимыми переменными (факторами) $F_1, F_2,...,F_m$. С целью выявления этой связи проводится серия экспериментов, в которой каждому опыту $(F_{j1}, F_{j2},...,F_{jm})$ ставится в соответствие его результат –значение зависимой переменной y_j, $j = 1, 2,....,n$. Искомая связь обычно описывается полиномом Колмогорова-Габора, который в простейшем случае имеет вид:

$$y_j = x_0 + F_{j1}x_1 + F_{j2}x_2 + ... + F_{jь}x_m + \varepsilon_j .$$

Здесь

F_{ji} - значение i-й независимой переменной в j-м опыте, $i = 0, 1, 2,...,m$, $j = 1, 2,...,n$;

y_j - значение объясняемой переменной в j-м опыте, $j = 1, 2,...,n$;

ε_j - значение случайной ошибки в j-м опыте, $j = 1, 2,...,n$.

В матричной форме приведенное соотношение имеет вид $FX = Y$, где

$$F = \begin{pmatrix} 1 & F_{11} & F_{12} & \dots & F_{1m} \\ 1 & F_{21} & F_{22} & \dots & F_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & F_{n1} & F_{n2} & \dots & F_{nm} \end{pmatrix}, \; X = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ \dots \\ x_m \end{pmatrix}, \; Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}.$$

Отыскание неизвестных коэффициентов X полинома (1) осуществляется методом наименьших квадратов следующим образом. С использованием (1) - (2) рассчитываются прогнозируемые значения объясняемой переменной $\hat{y}$ в каждом эксперименте, которые сравниваются с реальными значениями. Сумма квадратов получающихся отклонений – критерий адекватности искомого набора коэффициентов X. При этом формула для расчета вектора X имеет вид

$$X = \left(H^T H\right)^{-1} H^T Y.$$

Точность решения задачи зависит от правильности (адекватности) модели (1) и качества исходного статистического материала, сосредоточенного в матрице F. В свою очередь, адекватность модели определяется правильностью выбора объясняющих факторов (показателей) $\left(F_1, F_2, ..., F_m\right)$. В реальной задаче были отобраны следующие факторы, предположительно влияющие на результирующий показатель y - оставшаяся продолжительность жизни пациента, перенесшего операцию:

F_1 - возраст пациента;

F_2 - отдел желудка;

F_3 - G-дифференц.;

F_4 - инвазия в стенку желудка;

F_5 - VEGFR;

F_6 - p53;

F_7 - her2/new;

F_8 - нервы;

F_9 - сосуды;

F_{10} - инвазия в соседние органы;

F_{11} - объем лимфодиссекции;

F_{12} - л/у поражены;

F_{13} - MBP;

F_{14} - генетический тип.

Предварительный анализ результатов непосредственного оценивания значений факторов $\left(F_1, F_2, ..., F_{14}\right)$ позволил удалить из этого набора мало информативные факторы F_8, F_9, значения которых более чем в 95% случаев были одинаковыми и равными нулю.

Далее был приведен корреляционный анализ связи оставшихся факторов с результирующим показателем. По результатам этого анализа традиционно рекомендуется считать значимыми (влияющими) только те факторы, у которых коэффициент корреляции имеет значение, не ниже 0,15. Полученный набор коэффициентов корреляции сведен в таблицу 1.

Таблица 1

Коэффициенты корреляции между влияющими факторами и результирующей переменной y

Факторы	F_1	F_2	F_3	F_4	F_5	F_6
Коэффициент корреляции	**-0,15**	-0,03	0,09	-0,14	-0,03	0,08
Факторы	F_7	F_{10}	F_{11}	F_{12}	F_{13}	F_{14}
Коэффициент корреляции	0,01	-0,14	-0,04	-0,13	-0,12	-0,04

Из таблицы 1 видно, что минимально значимым является единственный коэффициент $k_{F_1,y}$, описывающий проявляющую себя зависимость продолжительности оставшейся жизни y от фактора F_1 (возраст). Наблюдаемый эффект отсутствия корреляции для остальных факторов, по-видимому, является следствием неконтролируемого и непрогнозируемого влияния на результат каких-либо неучтенных факторов (например, сопутствующие патологии, различия в характере и особенностях проживания в послеоперационный период и т.д.). После удаления факторов, значения которых заведомо не коррелируют со значениями результирующего фактора, был определен набор данных, пригодный для регрессионного анализа. Полученный в результате полином, по-прежнему, формирует неудовлетворительное по качеству предсказание значения оставшейся жизни. При этом рассчитываемые с помощью модели значения результирующего фактора существенно отличаются от наблюдаемых, что, по-видимому, также является следствием посторонних неконтролируемых влияний. Единственный и естественный путь снижения этого эффекта – группирование пациентов, имеющих одинаковые значения контролируемых показателей, и усреднение результирующего показателя y в пределах каждой группы. При этом все значения фактора F_1 (возраст) разделены на 10, а результат округлен до ближайшего целого, что обеспечило возможность группирования. В результате число групп оказалось равным 24. Для полученного сгруппированного набора данных вновь была проделан корреляционный анализ, который показал, что оставшиеся факторы значимо влияют на среднее в группе значение результирующего показателя.

Таблица 2

Коэффициенты корреляции между влияющими факторами и результирующей переменной y

Факторы	F_1	F_{10}	F_{13}
Коэффициент корреляции	-0,34	-0,40	-0,47

Таким образом, были получен окончательный набор исходных данных для регрессионного анализа.

Таблица 3.

Значения факторов и результирующего показателя для каждой группы

Возраст	Инвазия в соседние органы	МВР	Время жизни, мес.
6	1	1	40,15
5	0	0	52,22
7	1	1	35,29
4	1	1	36,00
5	1	1	49,25
8	1	1	33,80
6	2	1	20,00
6	3	1	45,00
5	3	1	38,33
6	0	0	49,90
7	1	0	43,33
7	0	0	48,40
4	0	0	55,17
9	1	1	14,50
8	1	0	31,00
8	0	0	40,00
4	1	0	48,00
7	2	1	31,50
3	0	0	46,00
6	1	0	43,00
5	4	0	26,00
6	2	0	31,80
7	3	1	28,00
8	0	0	41,00

В результате регрессионного анализа получен следующий полином

$$y = 58.16 - 2.31F_1 - 2.27F_{10} - 6.93F_{13}.$$

Проверку адекватности полученного уравнения регрессии проведем по критерию Фишера.

Вычислим величину общего рассеяния наблюдаемых значений результирующей переменной y относительно среднего их значения.

$$Q = \sum_{j=1}^{n}\left(y_j - \bar{y}\right)^2,\ \bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j .$$

Вычислим далее общее рассеяние наблюдаемых значений относительно значений, предсказываемых моделью

$$Q_{ост} = \sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2 .$$

Значение $Q_{ост}$ отражает влияние всех причин рассеяния, которые не может объяснить

введенное уравнение регрессии.

Наконец, вычислим параметр $F_{наблюдаемое}$, определяющий уровень адекватности полученного уравнения регрессии

$$F_{наблюдаемое} = \frac{(Q - Q_{ост})(n - m - 1)}{mQ_{ост}}.$$

В этом соотношении $(n - m - 1)$ и m - числа степеней свободы при расчете значений Q и $Q_{ост}$.

Вычисленное значение $F_{наблюдаемое}$ сравнивается с значением $F_{критическое}$, извлеченного из таблицы распределения Фишера с заданным уровнем значимости $\alpha = 0.05$ для заданного числа степеней свободы. Если при этом $F_{наблюдаемое} > F_{критическое}$, то гипотеза об адекватности уравнения регрессии (4) принимается, в противном случае эту гипотезу следует отклонить.

В рассматриваемой задаче

$Q = 2426{,}31$; $Q_{ост} = 1074{,}34$; $m = 3$; $n = 24$.

При этом $F_{наблюдаемое} = 8{,}39$. Значение $F_{критическое} = 5{,}93$.

Таким образом, полученная модель адекватна. Она позволяет с удовлетворительной точностью рассчитывать ожидаемое среднее значение продолжительности оставшейся жизни для заданного комплекса значений трех значимых показателей: возраст, инвазия в соседние органы и МВР.

Реальный путь улучшения качества прогноза – выявление значимых факторов, не учтенных в приведенном здесь исследовании, и увеличение объема статистического материала.

Литература/References:

1. STATISTICA 6. Статистический анализ данных. Второе издание. М.: Бином, 2009. Халафян А. А. ГЛАВА 14 Анализ выживаемости

2. A Prognostic Model to Predict Mortality among Non-Small-Cell Lung Cancer Patients in the U.S. Military Health System. Lin J, Carter CA, McGlynn KA, Zahm SH, Nations JA, Anderson WF, Shriver CD, Zhu K. J Thorac Oncol. 2015 Dec;10(12):1694-702. doi: 10.1097/JTO.0000000000000691. PMID:26473644.

3. Development of individual survival estimating program for cancer patients' management. Chang MC.Healthc Inform Res. 2015 Apr;21(2):134-7. doi: 10.4258/hir.2015.21.2.134. Epub 2015 Apr 30. PMID:25995966.

4. A new scoring system for predicting survival in patients with non-small cell lung cancer. Schild SE, Tan AD, Wampfler JA, Ross HJ, Yang P, Sloan JA.

5. Cancer Med. 2015 Sep;4(9):1334-43. doi: 10.1002/cam4.479. Epub 2015 Jun 23. PMID: 26108458.

Работа поступила в редакцию 28.11.2017 года.
Рекомендована к печати на заседании редакционной коллегии после рецензирования