УДК 616-07:621.3.012.3:04.438

А.А. Корнеенков, С.Г. Кузьмин,
В.Б. Дергачев, Д.Н. Борисов

Создание номограмм для оценки риска неблагоприятного клинического исхода

Военно-медицинская академия им. С.М. Кирова, Санкт-Петербург

***Резюме**: Представлена методика по разработке номограмм для оценки и стратификации риска определенного клинического исхода на основе созданного виртуального набора данных с использованием программной среды R. Виртуальный набор данных включал входные числовые и факторные переменные (типы переменных соответствуют документации программной среды R) и исход. Для количественных переменных были вычислены описательные статистики на всех уровнях переменной исхода, а для факторных переменных были построены мозаичные диаграммы. В качестве модели, описывающей ассоциацию входных переменных с исходом, была использована модель логистической регрессии. Для валидации и оценки производительности модели применен метод бутстрапа. Рассчитанные показатели валидности показали приемлемую дискриминационную способность прогностической модели. Статистическая калибровка продемонстрировала близость калибровочной кривой модели к идеальной калибровочной кривой. На основе коэффициентов логистической регрессии построена номограмма, по которой рассчитывалось значение риска определенного исхода для каждого оцениваемого субъекта (пациента). Показано, что с помощью представленной методики можно эффективно стратифицировать пациентов по риску неблагоприятного исхода, адекватно изменяя таким образом тактику диагностики и лечения. Использование номограммы значительно упрощает оценку риска и может быть использовано в бумажном виде в качестве дополнения к протоколу обследования пациента. В тексте статьи представлены коды языка программирования R с пояснениями.*

***Ключевые слова**: номограмма, виртуальные данные, метод бутстрапа, прогнозирование клинического исхода, логистическая регрессия, язык R, валидация модели, стратификация риска, оценка риска.*

Введение. Оценка и стратификация риска неблагоприятных исходов играют важную роль в клинической медицине. Создание простых в практическом использовании инструментов для поддержки принятия врачебных решений является важной задачей медицинской информатики и статистики. Благодаря достижениям в области информационных технологий и доступности медицинских данных из электронных историй болезни можно разработать эффективные оценочные модели оценки риска мало распространенных заболеваний в популяции. Как правило, в основе моделей риска лежит аддитивная модель прогноза, в которую факторы входят в виде алгебраической суммы, а их вклад в вероятность интересующего исхода представляется в виде весового коэффициента. Использование этих моделей требует применения компьютерной техники или калькуляторов, однако существуют способы значительно упростить их использование на практике, к которым относятся номограммы, шкалы, скоринговые карты и др.

Для создания практических инструментов прогнозирования риска нами использовался язык R, предназначенный для статистического анализа и управления данными. Язык R является языком программирования и программной средой с открытым исходным кодом, который доступен по адресу https://cran.r-project.org. Все функции, необходимые для разработки системы оценки, также доступны в составе соответствующих программных пакетов языка R. Синтаксис языка R хорошо документирован, что играет решающую роль при проверке решения задач. Предполагается, что пользователь рассматриваемой методики владеет базовыми знаниями и навыками использования языка R, описанными, например, в известной книге Р. Кабакова [3].

Цель исследования. Разработать номограммы для оценки риска определенного клинического исхода на основе виртуального набора данных с использованием R-языка.

Материалы и методы. Для иллюстрации применяемых методов использованы виртуальные данные, сгенерированные на основе опубликованных результатов клинических исследований факторов риска осложнений при оперативном лечении пациентов, страдающих болезнью Меньера [1]. Генерация виртуальной выборки обеспечивает воспроизводимость данных и решения всех задач.

Результаты и их обсуждение. Генерация исходных данных предусматривает использование программного кода R-языка (табл. 1), который создает набор данных, содержащий 5 переменных, включая бинарный (дихотомичный) результат «Исход» (outcome), две факторные переменные и две числовые переменные «Возраст» (age) и «СКУ» (sku) (сводный

Таблица 1

Программный код создания виртуальной выборки

Строка	Программный код R-языка
	library(dummies)
	set.seed(666)
	n<-150
	b0<-22
	age<-round(rnorm(n, mean=46.3, sd=10))
	sku<-round(rnorm(n, mean=65.6, sd=2.7),1)
	NYSTAGMUS<-as.factor(sample(x=c("no", "yes"),
	size=n,
	replace=TRUE,
	prob=c(0.42, 0.58)))
	VOMITUS<-as.factor(sample(x=c("no", "yes"),
	size=n,
	replace=TRUE,
	prob=c(0.47, 0.53)))
	VOMITUS<-relevel(VOMITUS, ref="yes")
	lp<-b0+0.12*age-0.41*sku+cbind(1, dummy(NYSTAGMUS)[, -1]) %*% c(0.001, 2.5)-cbind(1,dummy(VOMITUS)[,-1]) %*% c(0,2.2)+0.001*age^2-0.001*sku^2
	pi.x<-exp(lp)/(1+exp(lp))
	OUTCOME<-rbinom(n=n, size=1, prob=pi.x)
	df<-data.frame(OUTCOME, age, sku, NYSTAGMUS, VOMITUS)
	df$dataset<-sample(x=c("train", "validate"),
	size=n,
	replace=TRUE,
	prob=c(0.75, 0.25))
	f.y<-glm(OUTCOME~age+sku+NYSTAGMUS+VOMITUS, data=df, family="binomial")
	summary(f.y)

коэффициент устойчивости – СКУ) – суммарный показатель отклонения центра тяжести в ответ на различные динамические стимулы, дающий комплексное представление о состоянии статического и динамического равновесия. Переменная «Рвота» (vomitus) имеет два уровня: «no», «yes» (т. е. отсутствие признака и его наличие). Переменная «Спонтантанный нистагм» (nystagmus) также имеет два уровня: «no», «yes». Значение переменной «Исход» равное 1 соответствует наличию осложнения операции, 0 – его отсутствию.

Описание программы следующее. Строка 1 вызывает функцию «library» для загрузки пакета «dummies» в программу R. В строке 2 с помощью функции «set.seed» устанавливается произвольное значение переменной генератора случайных чисел, чтобы сделать результаты полностью воспроизводимыми. Строка 3 устанавливает размер выборки для примера (150 наблюдений). В строке 4 присваивается значение свободному члену регрессионного уравнения (b_0). Строки 5 и 6 генерируют количественные переменные: «Возраст» и «СКУ», которые распределены по нормальному закону распределения со средним значением 46,3 лет и 65,6%, а также стандартным отклонением 10 лет и 2,7% соответственно. В строке 7 создается факторная переменная – «Спонтанный нистагм» с двумя уровнями «no», «yes», количеством наблюдений n (строка 8), с заменой имеющихся значений на вставляемые (строка 9), и соответствующими пропорциями каждого из двух уровней (строка 10). В строках 11–14 таким же образом создается факторная переменная «Рвота». В строке 15 уровни этой переменной переопределяются, базовым уровнем назначался «yes». Строка 16 создает линейный предиктор «lp» уравнения логистической регрессии. Функция «dummy» используется для преобразования факторной переменной в фиктивные или «дамми» (dummy) переменные. Символом «%*%» обозначается матричный оператор умножения. Для каждой переменной задается линейный предиктор. Например, коэффициент 2,5, соответствующий уровню «yes» (строка 16) переменной «Спонтанный нистагм», интерпретируется как увеличение линейного предиктора на 2,5 единицы для «yes», по сравнению с «no». Строка 17 преобразует линейный предиктор в вероятность посредством логит-преобразования, а строка 18 генерирует переменную исхода, которая распределена по биномиальному распределению. Строка 19 объединяет все переменные в фрейм данных (data frame). Весь набор данных с помощью функции «sample» разделяется на подмножества данных для обучения и для валидации (строки 20–23). Три четверти всего набора сгенерированных данных используется для обучения модели (строка 23), а оставшаяся четверть используется для ее проверки. С помощью кода в строке 24 формируется обобщенная линейная модель «glm» под именем «f.m», по которой для общего контроля процесса с помощью функции «summary» выводятся коэффициенты и показатели созданной модели.

Перед созданием модели переменные были статистически описаны и оценены. Среди анализируемых в наборе данных 2 переменные: «Возраст» и «СКУ», являются числовыми переменными, выраженными в абсолютной шкале и 2 переменных «Спонтанный нистагм» и «Рвота» – факторными, выраженными в шкале наименований. Названия типов переменных соответствует документации R-языка.

Для представления результатов описательной статистики числовых переменных (табл. 2) приводятся команды языка R. В первой строке с помощью функции «library» указывается пакет «Hmisc», из которого используются некоторые функции. В строке 2 задается опция вывода отчета в формате LaTeX (формат вывода научных результатов, принятый в большинстве ведущих научных журналов). Во второй строке определяется перечень переменных, из которых в строке 4 формируется набор данных. В строках 5 и 6 переменные «Возраст» и «СКУ» определяются как числовые переменные и в строках 7 и 8 им присваиваются единицы измерения «лет» и «%» соответственно. Командой в строке 9 выводится отчет с описательными статистиками в программных кодах формата LaTeX. Чтобы преобразовать полученные коды в изображение или pdf-файл был использован один из многочисленных бесплатных LaTeX-редакторов и LaTeX-вьюверов – Papeeria (https://papeeria.com).

Таблица 2

Программный код описательной статистики числовых переменных модели

Строка	Программный код R-языка
	library(Hmisc)
	options(prType='latex')
	v<-c("age", "sku")
	df_descr<-df[, v]
	df_descr$age<-as.numeric(df_descr$age)
	df_descr$sku<-as.numeric(df_descr$sku)
	units(df_descr$age)<-"лет"
	units(df_descr$sku)<-"%"
	latex(describe(df_descr), file ="')

Описательная статистика переменных, включая квантили числовых переменных «Возраст» и «СКУ», представлена на скриншоте отчета-машинограммы (рис. 1), полученного при выполнении программного кода R-языка. В отчете используется несколько обозначений, которые требуют пояснения: «distinct» – указывает на число уникальных значений в наборе данных, «Info» – относительная мера информации, связана с тем, насколько непрерывна переменная. Наименьшее значение «Info» имеет переменная, имеющей только одно уникальное значение. «Gmd» – средняя разница или коэффициент Джини (англ., Gini's mean difference) – мера дисперсии, которая представляет собой среднюю абсолютную разницу между любыми парами наблюдений в наборе данных. «Lowest» и «highest» – соответственно минимальные и максимальные значения в наборе данных.

Графические методы для описания категориальных или факторных переменных и их ассоциации с исходом используются не так часто, как для количественных переменных. Метод логистической регрессии является аналогом дисперсионного и регрессионного анализа, в котором вместо количественного исхода моделируется категориальный, точнее, дихотомический исход. Графическое представление ассоциации между категориальными факторными переменными и дихотомическим исходом обычно ограничивается малопривлекательными диаграммами на основе таблиц сопряженности R×C (от англ., row – строка, column – столбец).

В качестве примера визуализации данных из таблиц сопряженности можно привести эффектный

Таблица 3

Программный код для создания мозаичной диаграммы ассоциации спонтанного нистагма и рвоты с переменной исхода

Строка	Программный код R-языка
	libraryvcd)
	mytable<-xtabs(~OUTCOME+NYSTAGMUS+VOMITUS, data=df)
	mosaic(mytable, shade=TRUE, legend=TRUE,
	set_labels=list(VOMITUS = c("no", "yes"),
	NYSTAGMUS=c("no", "yes"),
	OUTCOME=c("good", "bad")),
	pop=FALSE)
	labs<-round(prop.table(mytable), 2)
	labeling_cells(text=labs, margin=0)(mytable)

метод построения и анализа мозаичных диаграмм. Мозаичные диаграммы, предложенные J.A. Hartigan, B. Kleiner [6], представляют каждую ячейку таблицы сопряженности прямоугольником (или плиткой), площадь которого пропорциональна частоте ячейки. Рисунок из плиток на мозаичной диаграмме полезен для выдвижения статистических гипотез, визуального сравнения частот в таблице сопряженности и выделения среди них необычно больших и малых значений.

Для решения нашей задачи визуализации ассоциации факторных переменных: «Спонтанный нистагм» и «Рвота» с бинарным исходом была использована мозаичная диаграмма. В первой строке кода (табл. 3) подключается пакет «vcd», с помощью которого реализуется построение мозаичной диаграммы. В строке 2 формируется таблица «mytable» с факторными переменными и результирующей переменной исхода из фрейма данных «df». В строках 3–7 создается мозаичная диаграмма с метками категорий переменных. В строках 8 и 9 формируются метки в центре плиток мозаичной диаграммы с округленными до двух знаков после запятой частотами.

На рисунке 2 представлен результат формирования мозаичной диаграммы для представления ассоциации трех переменных модели: «Исход», «Спонтанный нистагм» и «Рвота».

Цвета плитки представляют величину остатков (разности наблюдаемых и ожидаемых частот) для ячейки или комбинации уровней. Синий цвет плитки означает, что в этой ячейке больше наблюдений, чем

Возраст, лет

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
150	0	41	0.999	45.58	12.08	29.0	33.0	38.0	46.0	53.0	59.0	64.1

lowest : 15 17 24 26 28, highest: 65 66 67 68 72

СКУ, %

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
150	0	84	1	65.34	3.22	60.64	61.59	63.70	65.35	67.47	69.10	69.70

lowest : 58,1 58,5 58,7 59,1 59,7, highest: 70,2 70,5 71,0 71,1 71,5

Рис. 1. Скриншот отчета в формате LaTeX с описательной статистикой числовых переменных

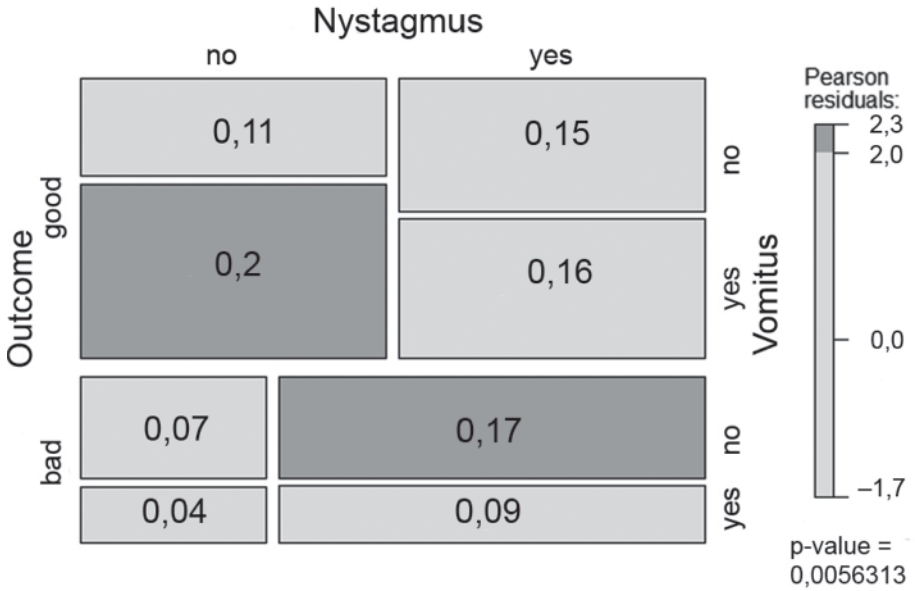

Рис. 2. Мозаичная диаграмма ассоциации переменных спонтанного нистагма, рвоты и исхода

можно было бы ожидать при нулевой модели (модель, соответствующая нулевой гипотезе H_0 – независимости переменных). Красный цвет означает, что наблюдений меньше, чем можно было бы ожидать. Цвета, тестовая статистика остатков, значение p в легенде справа от мозаичной диаграммы показывают, какие ячейки вносят статистически значимый вклад в исход по тесту хи-квадрат.

Более сложным методом визуализации нелинейной ассоциации предикторов и вероятности исхода является метод сглаживания «loess». Метод «loess» или метод локальных полиномиальных регрессий (от англ. LOcal regrESSions – «LOESS» или LOcally WEighted Scatterplot Smoother – «LOWESS») в настоящее время широко используется в прогнозировании, исследовании различных тенденций [2] и подробно описан в работах W.S. Cleveland [4]. Метод позволяет сгладить ряд значений, используя простую линейную либо полиномиальную зависимость двух переменных y и x. При этом предлагается строить модель не по всему ряду данных, а по его отдельным частям или диапазонам. Идея метода заключается в том, чтобы рассчитать множество регрессионных моделей, центрами каждой из которых поочередно являются значения из ряда данных. Поскольку это непараметрический метод, он не требует предположения о параметрах распределения.

В таблице 4 представлен программный код R-языка для построения четырех loess-диаграмм с использованием пакета «ggplot2». В строках 3–6 формируются четыре диаграммы p1, p2, p3, p4 на основе взаимодействия разных переменных. Чтобы можно было различать линии разных переменных, им присваиваются разные характеристики, для переменной «Рвота» – цвет, для переменной «Спонтанный нистагм» – размер линии и на четвертой диаграмме их сочетание.

Таблица 4

Программный код построения loess-диаграмм

Строка	Программный код R-языка
	library(ggplot2)
	yl<-ylab(NULL)
	p1<-ggplot(df, aes(x=sku, y=OUTCOME))+histSpikeg(OUTCOME~sku, lowess=TRUE, data=df)+ylim(0,1)+yl+ylab("Probability of Bad Outcome")
	p2<-ggplot(df, aes(x=sku, y=OUTCOME, color=VOMITUS))+histSpikeg(OUTCOME~sku+VOMITUS, lowess=TRUE, data=df)+ylim(0,1)+yl
	p3<-ggplot(df, aes(x=sku, y=OUTCOME, size=NYSTAGMUS))+histSpikeg(OUTCOME~sku+NYSTAGMUS, lowess=TRUE, data =df)+b+ylim(0,1)+yl+ylab("Probability of Bad Outcome")
	p4<-ggplot(df, aes(x=sku, y=OUTCOME, color=VOMITUS, size=NYSTAGMUS))+histSpikeg(OUTCOME~sku+VOMITUS+NYSTAGMUS, lowess=TRUE, data=df)+b+ylim(0,1)+yl
	gridExtra::grid.arrange(p1,p2,p3,p4,ncol=2)

В строке 7 производится объединение этих диаграмм в две колонки (ncol=2) на одном общем рисунке (рис. 3).

Все представленные методы позволяют одним из многочисленных способов провести предварительный анализ факторного пространства, определить ассоциации переменных, сформулировать статистические гипотезы и т. д. Следующим этапом разработки номограммы является создание логистической регрессионной модели. Чем качественнее и точнее логистическая регрессионная модель, тем качественнее и точнее предсказывающая способность номограммы.

Цель логистической регрессии – найти наилучшим образом подогнанную (но все же биологически разумную) модель, описывающую взаимосвязь между интересующей дихотомической результирующей

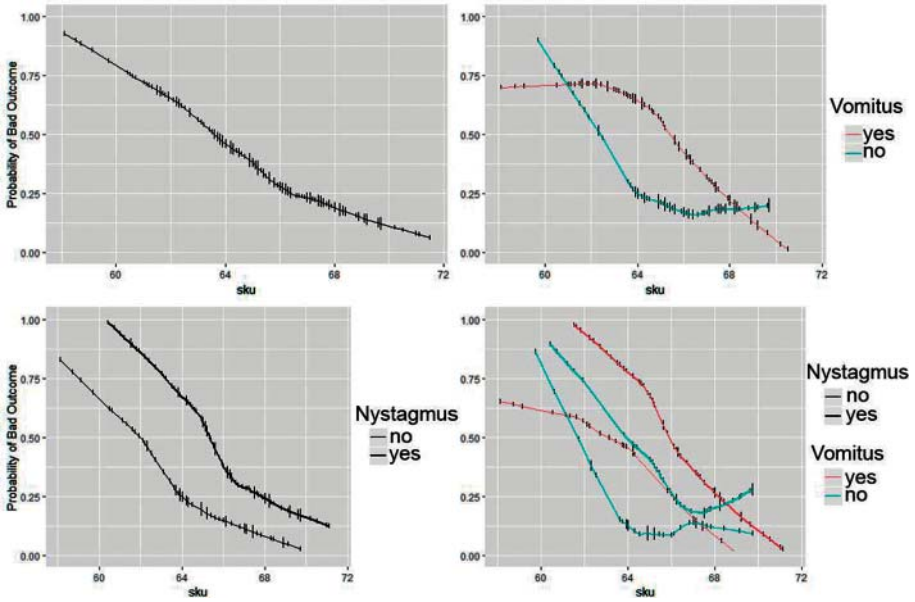

Рис. 3. Loess-диаграмма зависимости вероятности неблагоприятного (bad) исхода от СКУ и от меняющихся на фоне СКУ значений факторных переменных спонтанного нистагма и рвоты

переменной и набором независимых факторных переменных (предикторов). Логистическая регрессия формирует коэффициенты регрессии (их стандартные ошибки и уровни значимости) по формуле логит-преобразования вероятности интересующего значения результирующей переменной:

$$\text{logit}(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_k,$$

где p – вероятность наличия признака, представляющего интерес, b – коэффициенты модели, X – значения предикторов. Логит-преобразование определяются как логарифм шансов:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right).$$

Модель может использоваться для расчета предсказанной вероятности исхода (p) для заданных значений предикторов:

$$p = \frac{e^{(\text{logit}(p))}}{1 + e^{(\text{logit}(p))}}.$$

После оценки коэффициентов модели существует несколько шагов, связанных с оценкой адекватности и полезности модели. Вначале важность каждой из объясняющих переменных оценивается путем проведения статистических тестов значимости коэффициентов. Затем проверяют общую пригодность модели. Кроме того, оценивается способность модели различать две группы, определенные результирующей переменной исходом. Наконец, если это возможно, модель подтверждается проверкой валидности и дискриминационной способности на отличном от того набора данных, который использовался для разработки и обучения модели.

В таблице 5 представлены команды для создания диаграммы (рис. 4), отражающей значимость предикторов по хи-квадрату. В строке 1 представлен код для создания модели логистической регрессии с помощью функции «lrm» (от англ., logistic regression model).

Таблица 5

Программный код для рисунка с ранжированием важности предикторов

Строка	Программный код R-языка
	f<-lrm(formula=OUTCOME~age+sku+VOMITUS+NYSTAGMUS, data=df)
	an<-anova(f)
	plot(an)

В строке 2 проводится дисперсионный анализ созданной модели, в строке 3 – графический вывод его результатов.

Для вывода результатов моделирования в формате LaTeX используется функция «latex». Вероятность исхода, Prob{Outcome=1}, рассчитывается по формуле:

$$\text{Prob}\{\text{OUTCOME} = 1\} = \frac{\exp(-X\hat{\beta})}{1 - \exp(-X\hat{\beta})},$$

$$X\hat{\beta} = 31{,}3 + 0.24\,\text{Age} - 0.68\,\text{Sku} - 1.44\,\text{Vomitus [no]} + 3{,}32\,\text{Nystagmus [yes]},$$

где $X\hat{\beta}$ – линейный предиктор модели.

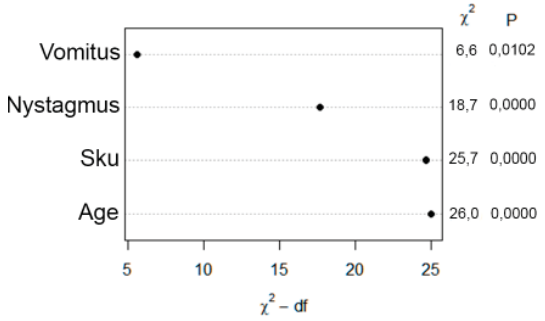

Рис. 4. Ранжированная по возрастанию важность ($\chi^2 - df$) предикторов модели

Валидация модели включает в себя оценку показателей дискриминационной способности модели и калибровку. Для получения этих показателей исходные данные специальным образом подготавливаются с помощью метода бутстрапа (bootstrap methods), который в настоящее время является методом выбора для решения подобных задач [5]. Суть метода заключается в том, что из имеющегося набора данных случайным образом отбирается заданное число выборок, по каждой из которых рассчитываются необходимые показатели дискриминационной способности и калибровки. Эти показатели используются для корректировки показателей, рассчитанных на исходном наборе данных (оригинальной выборке), таблица 6.

Таблица 6

Программный код для вывода результатов валидации модели

Строка	Программный код R-языка
1	cal<-calibrate(f, B=200)
2	plot (cal)
3	options(prType='latex')
4	v <- validate (f , B =200)
5	print (vfull , digits =3)

Применительно к нашей задаче, в начале моделирования исходный набор данных (original sample) был случайным образом в соотношении 75:25 разделен на две выборки: обучающую (training sample) и пробную (test sample) выборки. Считается, что модель по обучающей выборке имеет лучшие значения показателей производительности, поэтому она обладает определенным «оптимизмом» по сравнению с моделью, полученной на пробной выборке. Разница показателей производительности для обучающей и пробной выборок называется «оптимизмом» (optimism) модели, для получения скорректированного показателя (corrected index) значение «оптимизма» вычитается из показателя, полученного на исходном наборе данных.

Прогнозируемая вероятность исхода (Predicted Pr{Outcome=1}) строится на фоне наблюдаемой вероятности (Actual Probability), а отклонение от идеальной линии указывает на разницу между прогнозируемыми и наблюдаемыми рисками (рис. 5).

Близость калибровочной кривой к диагональной линии 45° демонстрирует приемлемую валидацию по шкале абсолютной вероятности. В таблице 7 представлены показатели (индексы) дискриминационной способности полученной модели. Например, D_{xy} – ранговая корреляция Сомерса (Somers' D) между прогнозируемой вероятностью того, что переменная исхода = 1, и наблюдаемыми значениями переменной исхода; D – показатель дискриминации, отношение правдоподобия модели χ^2, деленное на размер выборки; R^2 – псевдо R-квадрат (R^2 Nagelkerke-Cox-Snell-Maddala-Magee), коэффициент детерминации модели. Более подробную информацию о представленных валидационных показателях можно почерпнуть из описания пакета «rms» в книге F.E. Harrell [7]. Основные показатели демонстрируют приемлемую оценку валидации модели (табл. 7).

Конечным методом отображения взаимосвязи между несколькими предикторами и вероятностью ответа является построение номограммы. В таблице 8 представлены программные коды для построения номограммы.

На рисунке 6 показана номограмма, порядок использования которой можно проиллюстрировать следующим примером. Обследуется пациент в возрасте 45 лет, показатель СКУ=65%, рвота не наблюдается (Vomitus=«no»), спонтанный нистагм есть (Nystagmus=«yes»).

Для возраста 45 лет находится соответствие на шкале «Points» = 50 баллов. СКУ равное 65% соответствует примерно 33 баллам, отсутствие рвоты (vomitus= «no») равно 0 баллам, наличие спонтанного нистагма (Nystagmus= «yes») дает примерно 24 балла. В сумме 50+33+0+24=107 баллов соответствуют значению линейного предиктора около 0 и вероятности осложнения (т. е. Outcome =1) около 50%.

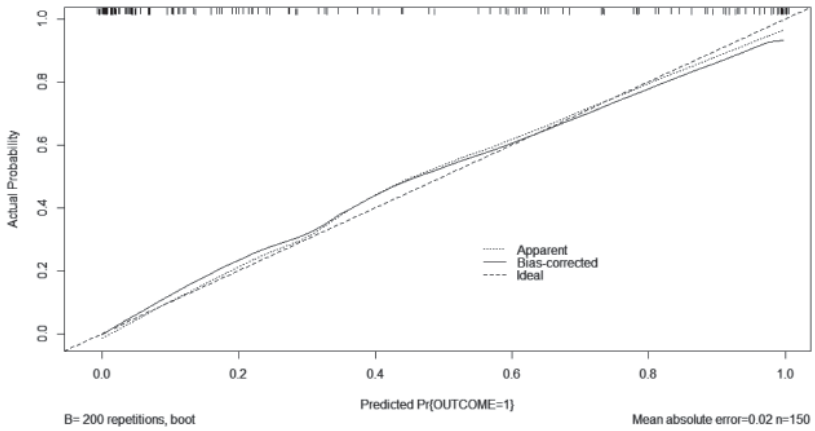

Рис. 5. Калибровочная кривая (КК) для модели исхода (Apparent – КК по оригинальному набору данных, Bias-corrected – КК, скорректированная на оптимизм, Ideal – идеальная КК)

Таблица 7

Статистика валидации модели

Показатель	Для исходного набора данных	Для обучающей выборки	Для пробной выборки	Значение «оптимизма»	Скорректированный показатель	n
D_{xy}	0,878	0,884	0,867	0,017	0,861	200
R^2	0,697	0,711	0,681	0,030	0,668	200
Intercept	0,000	0,000	–0,005	0,005	–0,005	200
Slope	1,000	1,000	0,889	0,111	0,889	200
E_{max}	0,000	0,000	0,027	0,027	0,027	200
D	0,709	0,731	0,685	0,046	0,663	200
U	–0,013	–0,013	0,007	–0,020	0,007	200
Q	0,723	0,715	0,678	0,067	0,656	200
B	0,098	0,093	0,104	–0,011	0,109	200
g	1,015	1,393	3,822	0,571	3,475	200
g_p	0,413	0,414	0,408	0,006	0,407	200

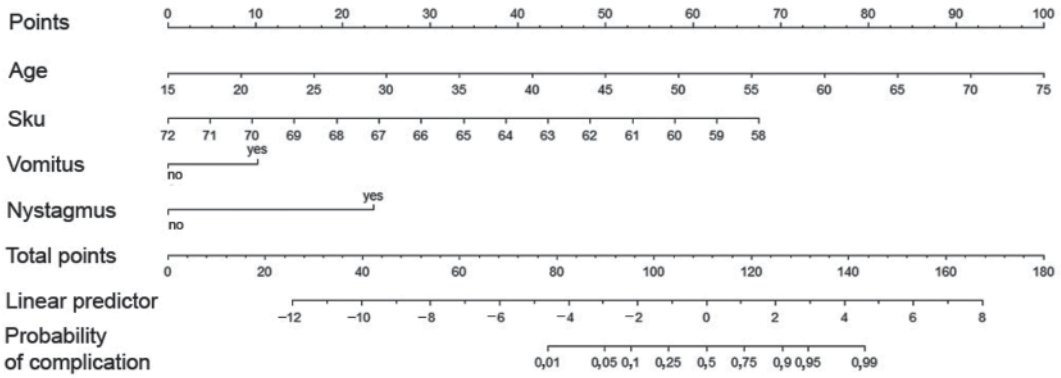

Рис. 6. Номограмма, вычисляющая $x\hat{\beta}$ (Linear predictor – линейный предиктор) и $\hat{P}$ для осложнения. Для каждого предиктора определяются соответствующие значения баллов на шкале 0–100, которые затем суммируются. Результат считывается по шкале «Total points», а затем вероятность осложнения считывается на нижней шкале «Probability of compilation»

Таблица 8

Программные коды для рисования номограммы

Строка	Программный код R-языка
1	nom <- nomogram (f, fun = plogis, funlabel ="Probability of comlication",
2	fun.at =c(.01, .05, .1, .25, .5, .75, .9,.95, .99))
3	plot (nom, xfrac = .45)

Заключение. Представленная методика построения номограммы с помощью R-языка позволяет улучшить понимание всего процесса создания этого эффективного диагностического инструмента. Подобные диагностические инструменты относятся к скрининговым методам диагностики, они позволяют стратифицировать пациентов по риску неблагоприятных исходов, увеличивая таким образом шанс пациента получить именно то лечение, которое будет для него наиболее результативным и наименее опасным.

Кроме того, построение номограмм, как и скоринговых карт, калькуляторов риска является хорошим способом отчуждения знаний и опыта экспертов для их использования молодыми специалистами, которые таким опытом и знаниями не обладают. Современные достижения медицинской информатики и статистики значительно упрощают этот процесс, обеспечивая его прозрачность и воспроизводимость.

Литература

1. Корнеенков, А.А. Использование модифицированной процедуры последовательного распознавания Вальда для определения исхода оперативного лечения у пациентов с болезнью Меньера / А.А. Корнеенков [и др.] // Росс. оториноларингол. – 2018. – № 3 (94). – С. 54–59.
2. Светуньков, И.С. Методы социально-экономического прогнозирования в 2 т. Т. 2 Модели и методы: учебник и практикум для академического бакалавриата / И.С. Светуньков, С.Г. Светуньков. – М.: Издательство Юрайт, 2018. – С. 17–25.
3. Кабаков, Р.И. R в действии. Анализ и визуализация данных в программе R / И. Кабаков; пер. с англ. П.А. Волковой. – М.: ДМК Прес, 2016. – 588 с.
4. Cleveland, W.S. Robust Locally Weighted Regression and Smoothing Scatterplots / W.S. Cleveland // American Statistical Association. – 1979. – Vol. 74, № 368. – P. 829–836.

5. Efron, B. Bootstrap methods: Another look at the jackknife / B. Efron // Ann. Statist. – 1979. – № 7. – P. 1–26.

6. Hartigan, J.A. Mosaics for Contingency Tables / J.A. Hartigan, B. Kleiner // Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface. – New York: Springer, 1981. – P. 268–273.

7. Harrell, F.E.Jr. Regression Modeling Strategies / E.F.Jr. Harrell. – Switzerland: Springer International Publishing, 2015. – 582 p.

A.A. Korneenkov, S.G. Kuzmin, V.B. Dergachev, D.N. Borisov

Development of nomograms to assess the risk of clinical outcome

Abstract. *A methodology is presented for developing nomograms for assessing and stratifying the risk of a clinical outcome based on the created virtual data set using the R software environment. The virtual data set included input numerical and factor variables (variable types correspond to the R software documentation) and outcome. For quantitative variables, descriptive statistics were calculated at all levels of the outcome variable, and mosaic diagrams were constructed for factor variables. As a model that describes the association of input variables with the outcome, a logistic regression model was used. A bootstrap method was applied to validate and evaluate the model performance. The calculated validity indicators showed an acceptable discriminatory ability of the predictive model. The statistical calibration demonstrated the proximity of the model's calibration curve to the ideal calibration curve. Based on the logistic regression coefficients, a nomogram was constructed using which the risk value of a specific outcome was calculated for each subject (patient). It is shown that with the help of the presented technique it is possible to stratify patients effectively by the risk of an adverse outcome, thus adequately altering the diagnosis and treatment tactics. The use of a nomogram greatly simplifies risk assessment and can be used in paper form as a supplement to the patient examination protocol. The article contains the codes of the R programming language with explanations.*

Key words: *nomogram, virtual data, bootstrap method, prediction of clinical outcome, logistic regression, R language, model validation, risk stratification, risk assessment.*

Контактный телефон: +7-904-554-07-40; email: vmeda-nio@mil.ru