ЛУЧИНИН А.С.,
к.м.н., ФГБУН КНИИГиПК ФМБА России, г. Киров, Россия,
e-mail: @niigpk.ruluchinin@niigpk.ru

ЛЯНГУЗОВ А.В.,
к.м.н., ФГБУН КНИИГиПК ФМБА России, г. Киров, Россия,
e-mail: lyanguzov@niigpk.rulyanguzov@niigpk.ru

ВЫБОР ПРЕДИКТОРОВ ДЛЯ МОДЕЛЕЙ КЛАССИФИКАЦИИ И ПРОГНОЗА В МЕДИЦИНЕ

DOI: 10.25881/18110193_2022_3_54

Аннотация.
Процессы создания максимально простой и одновременно эффективной прогностической модели в медицине должны быть сбалансированы. Факторы, включенные в модель, являются основой ее качества и практической значимости, однако их выбор — не всегда простая задача. Цель исследования — сравнение разных методов селекции предикторов для создания медицинских прогностических моделей.

Материалы и методы. Для выбора предикторов использовали такие методы, как корреляция, фильтрация признаков на основе базовой статистики, однофакторный анализ Хосмера-Лемешоу, так и сложные, которые часто используются в машинном обучении: рекурсивное исключение признаков, регрессия «LASSO» и деревья классификации. Прогностические модели построили с использованием метода бинарной множественной логистической регрессии. Статистический анализ проводился с использованием языка программирования R (версия 3.4.2).

Результаты. Наборы предикторов, полученные при помощи методов «LASSO» и случайного леса, а также методом пошаговой регрессии, позволили построить наиболее точные прогностические модели (минимальное значение AIC). Базовые методы статистического анализа и однофакторный регрессионный анализ по методу Хосмера-Лемешоу оказались наименее эффективными.
Заключение. Применение методов селекции предикторов часто существенно сокращает их количество, отсеивая неинформативные, что улучшает качество будущей модели прогноза.

Ключевые слова: *прогноз, прогностические модели, селекция предикторов.*

Для цитирования: *Лучинин А.С., Лянгузов А.В. Выбор предикторов для моделей классификации и прогноза в медицине. Врач и информационные технологии. 2022; 3: 54-67. doi: 10.25881/18110193_2022_3_54.*

LUCHININ A.S.,
PhD, The Federal State-Financed Scientific Institution Kirov Research Institute of Hematology and Blood Transfusion under the Federal Medical Biological Agency, Kirov, Russia, e-mail: luchinin@niigpk.ru

LYANGUZOV A.V.,
PhD, The Federal State-Financed Scientific Institution Kirov Research Institute of Hematology and Blood Transfusion under the Federal Medical Biological Agency, Kirov, Russia, e-mail: lyanguzov@niigpk.ru

FEATURE SELECTION FOR MEDICAL PROGNOSTIC MODELS

DOI: 10.25881/18110193_2022_3_54

Abstract.
It is very important to balance the processes of creating the simplest and most effective predictive models in medicine. The predictors in the model determine its quality and practical relevance but selecting them is not always easy. The aim of the study is to compare different methods of prediction selection to create medical prognostic models.

Methods. We compare simple methods, such as correlation, predictor filtering based on basic statistics, and Hosmer-Lemeshow univariate analysis, with more complex methods often used in machine learning, such as recursive feature elimination, LASSO regression, and classification trees. The predictive models were built using the binary multiple logistic regression method. Statistical analysis was carried out using the programming language R (version 3.4.2).

Results. Based on the LASSO and random forest methods, as well as the stepwise regression method, the most accurate predictive models were constructed (minimum AIC value). The Hosmer-Lemeshow method and basic methods of statistical analysis have been found to be the least effective.

Conclusion. The use of predictor selection methods often significantly reduces their number, filtering out non-informative ones, which improves the quality of the predictive model.

Keywords: *prognosis, predictive models, predictor selection*

For citation: *Luchinin A.S., Lyanguzov A.V. Feature selection for medical prognostic models. Medical doctor and information technology. 2022; 3: 54-67. doi: 10.25881/18110193_2022_3_54.*

ВВЕДЕНИЕ

Популярность прогнозной аналитики в медицинских исследованиях постоянно растет. Часто целью научных экспериментов является прогнозирование выживаемости пациентов, риска осложнений, эффективности лечения и др., а также создание различных классификаций [1–3]. Для этого формируются наборы данных, которые содержат результаты изучаемых событий и факторы, способные предсказать их наступление с той или иной точностью, — предикторы. Обычно исследователь выбирает определенный метод статистического анализа или машинного обучения, делает расчеты и интерпретирует полученные результаты, не всегда уделяя должного внимания выбору предикторов. Тем не менее, этот этап является крайне важным, так как влияет на точность всей будущей модели.

Существуют две основные проблемы анализа данных, с которыми сталкивается исследователь при прогнозировании событий и построении классификаций. Первая связана с большими трудозатратами, требующимися для получения и структурирования медицинских данных, вследствие чего имеет место дефицит информации и оцениваются выборки малого объема [4; 5]. Вторая — большое количество переменных или атрибутов, которые по мнению исследователя влияют на исход и должны быть проверены в ходе научного эксперимента [6]. Большой набор признаков может создавать избыточность информации — «шум», который нивелирует значимость важных факторов. Кроме этого, наличие большого числа переменных требует увеличения размера выборки. Оценка данных без их предварительной обработки может ухудшить результаты, так как наличие неинформативных переменных добавляет неопределенности и снижает общую эффективность прогнозирования. При использовании методов машинного обучения часто приходится иметь дело с большим количеством переменных. В этом случае обилие признаков также усложняет прогнозное моделирование, обусловливая явление, известное как «проклятие размерности» [7].

В связи с этим необходимо сокращение исходного избыточного числа входных данных без потери качества модели. Для этого используется селекция предикторов (СП) — метод подготовки данных, применяемый к ним перед моделированием [7]. Он может быть выполнен после очистки и масштабирования данных перед обучением прогностической модели.

Для СП разработаны специальные методы выбора и извлечения признаков, позволяющие повысить общий потенциал классификатора или прогностической модели [8]. Посредством СП происходят уменьшение вычислительных затрат на моделирование, упрощение модели, улучшение её качества и производительности, а также удаление неинформативных или избыточных предикторов.

Цель работы — описать различные методы СП и провести их сравнительный анализ на примере данных о летальных исходах пациентов с заболеваниями системы крови в отделении интенсивной терапии (ОИТ).

МАТЕРИАЛЫ И МЕТОДЫ

Исходные данные

Для прогнозного моделирования использованы клинические и лабораторные данные 202 пациентов, госпитализированных в ОИТ в связи с осложнениями основного гематологического заболевания. Возраст больных колебался от 19 до 82 лет (медиана — 57 лет), из них — 112 (55%) мужчин и 90 (45%) женщин. За исход принимали витальный статус пациента к концу госпитализации, который являлся бинарной переменной (1 — умер, 0 — жив). Все клинические случаи описали с использованием 21 предиктора (табл. 1). Значения предикторов получили в интервале ±2 суток от даты поступления в ОИТ.

Методы селекции предикторов

Все методы СП можно разделить на контролируемые (с учителем) и неконтролируемые (без учителя). Неконтролируемые методы выбора признаков игнорируют целевую (зависимую) переменную, например, используя для этой цели корреляцию. При контролируемой СП выбор признаков зависит от целевой переменной или ориентируется на ее меняющийся прогноз (Рис. 1).

В отдельную группу входят методы снижения размерности — преобразование данных из многомерного пространства в низкоразмерное (например, в двумерное) так, чтобы значимые

Таблица 1 — Характеристика предикторов

Предиктор	Значение*
Пол	Мужчины — 112 (55%) Женщины — 90 (45%)
Возраст, лет	57 (19–82)
Температура тела, C^0	37,0 (34,0–39,4)
ЧД, $мин^{-1}$	22 (12–50)
Систолическое АД, мм рт. ст.	115 (51–175)
Диастолическое АД, мм рт. ст.	70 (25–100)
Среднее АД, мм рт. ст.	86 (41–123)
ЧСС, $мин^{-1}$	95 (55–170)
Инотропная поддержка катехоламинами (Да/Нет)	21 (10%)
Гипоксемия (Да/Нет)	85 (42%)
Уровень сознания по шкале Глазго <15 баллов	26 (13%)
Гемоглобин, г/л	81 (74–94)
Лейкоциты, 10^9/л	1,2 (0–703)
Тромбоциты, 10^9/л	31 (1–807)
Креатинин, мкмоль/л	80 (36–471)
С-реактивный белок, г/л	0,116 (0–0,682)
Общий белок, г/л	59,5 (34,5–137)
Альбумин, г/л	33 (16,1–53,7)
Общий билирубин, мкмоль/л	12,6 (2,9–280,2)
Прокальцитонин, нг/мл	0,459 (0,019–125,52)
Бактериемия (Да/Нет)	40 (20%)

Примечание: * — количество пациентов (%) для категориальных переменных; медиана (минимум, максимум) для количественных переменных. ЧД — частота дыхания, АД — артериальное давление, ЧСС — частота сердечных сокращений.

Рисунок 1 — Классификация методов селекции предикторов с примерами.

присутствующие в них свойства не терялись при преобразовании [9]. Уменьшение размерности обычно используется при визуализации данных для их лучшего понимания и интерпретации исследователем, а также в методах машинного или глубокого обучения для упрощения поставленной задачи. Эти методы не рассматриваются в текущей работе.

Корреляция признаков лежит в основе проблемы мультиколлинеарности. Коррелирующие между собой предикторы, оказывающие влияние на результат, не позволяют однозначно оценить параметры исходной линейной регрессионной модели и правильно разделить вклад регрессоров в прогноз зависимой переменной. Иногда это приводит к логическим противоречиям (парадоксам), когда заведомо неблагоприятный предиктор становится фактором благоприятного прогноза [10; 11].

Использование оберточных методов СП позволяет оценить эффективность выбора подмножества признаков, учитывая финальный результат примененного алгоритма обучения, в частности по уровню прироста точности модели. Примером такого подхода является рекурсивное удаление признаков или RFE (Recursive Feature Elimination) — алгоритм поиска оптимального решения задачи, который оставляет только те признаки, которые вносят хоть какой-то вклад в модель, а все остальные исключаются [7]. При реализации метода RFE модель обучается на исходном наборе признаков и оценивает их значимость. Затем исключается один или несколько наименее важных предикторов, а модель обучается на оставшихся, и так далее, пока не останется определенное число наилучших признаков. Иными словами, алгоритм эффективно выбирает функции (столбцы) в обучающем наборе данных, которые наиболее релевантны для прогнозирования целевой переменной.

Пошаговая регрессия использует разные варианты СП. Прямой отбор (Forward stepwise) — итеративное добавление признаков к изначально пустому набору с целью наилучшего прироста качества модели. Обратный отбор (Backward stepwise) — итеративное удаление предикторов, начиная с набора, состоящего из всех признаков, обеспечивающее наилучший прирост качества модели [12].

К группе встроенных методов СП относятся алгоритмы, которые одновременно обучают модель и отбирают признаки, например регуляризация, представленная регрессией «LASSO» (Least Absolute Shrinkage and Selection Operator), деревья решений, алгоритм «случайного леса» и др. Регуляризация — стратегия, направленная на снижение переобучения, которая помогает создать более простые и точные модели [13]. Другим высокоэффективным методом СП является алгоритм «Boruta», названный так в честь лесного духа из польской и славянской мифологии и построенный на базе метода машинного обучения «случайный лес» (Random Forest). Последний является алгоритмом контролируемого ансамблевого обучения для классификации и регрессии путем голосования большого числа более слабых классификаторов — деревьев решений, которые создаются независимо друг от друга на разных подвыборках тренировочного набора данных [14]. В основе алгоритма Boruta лежит использование так называемых «теневых предикторов» — искусственно сгенерированных аналогов независимых переменных, значения которых случайным образом распределяются по отношению к целевой переменной. Каждый исходный предиктор сравнивается с порогом, за который принимается самая высокая прогностическая значимость, зарегистрированная среди «теневых» признаков. Идея заключается в том, что предиктор является полезным только в том случае, если он способен прогнозировать исход эффективнее, чем лучшая случайная функция. Найденные значимые предикторы ранжируются в порядке своей эффективности [15].

Фильтры — метод СП, при котором важность признаков определяется только на основе свойственных им характеристик, без привлечения алгоритмов обучения. Наиболее простыми из них являются методы базовой статистики для сравнения распределения частот в группах, такие как $\chi 2$ и точный критерий Фишера. Последний обычно применяется в случае, когда категориальные переменные имеют ожидаемую частоту 5 и менее наблюдений. В случае количественных переменных применяются методы оценки дисперсии, например ANOVA, или их непараметрические аналоги. Принцип СП при использовании фильтрационных методов

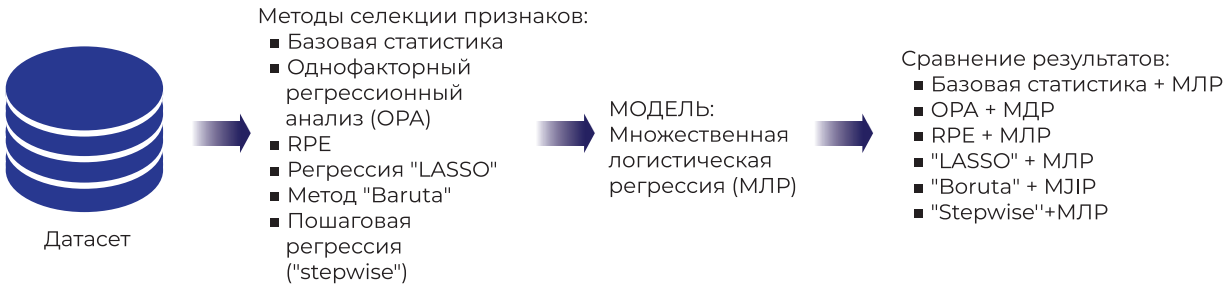

Рисунок 2 — Дизайн исследования.

строится на определении потенциальной прогностической значимости факторов, если распределение их значений (дисперсии или частоты) меняется в зависимости от исхода зависимой переменной.

Популярным методом СП является однофакторный анализ по методу Хосмера-Лемешоу, который может применяться в логистической регрессии и в модели пропорциональных рисков Кокса при анализе выживаемости. Все предикторы поочередно проверяются на предмет прогностической эффективности, после чего для многофакторного анализа выбираются только те, что показали статистическую значимость ниже порогового уровня альфа от 0,05 до 0,25 по рекомендациям разных исследователей [16]. Существует также гибридный способ отбора признаков, состоящий из комбинации разных методов.

Для СП в настоящей работе использованы все описанные выше методы, дизайн исследования представлен на рис. 2. В качестве прогностической модели с учетом бинарного исхода зависимой переменной (смерть в течение периода госпитализации в ОИТ) применили биномиальную логистическую регрессию. Так как целью исследования было сравнение различных методов СП, а не создание максимально точной модели прогноза летальности в ОИТ, какая-либо модификация предикторов (стандартизация или логарифмизация) перед моделированием не проводилась, тестовая выборка не создавалась. В качестве показателя точности модели использовали информационный критерий Акаике (AIC), который применяется для выбора лучшей статистической модели из нескольких, построенных на одном и том же наборе данных — чем меньше значение AIC, тем лучше модель [17].

Статистический анализ проводился на базе языка программирования R (версия 4.1.2), библиотеки: «missForest», «boruta», «glmnet», «caret».

Исходная матрица данных из 202 наблюдений и 21 предиктора содержала 4242 значения, из них 103 (2,4%) оказались пропущенными. Проблему отсутствующих значений решили путем вменения на их место синтетических параметров, спрогнозированных с помощью модели машинного обучения — «случайный лес» (библиотека R «missForest»).

РЕЗУЛЬТАТЫ

Корреляция

Корреляционный анализ применили для селекции количественных переменных. Коллинеарными считались переменные, у которых абсолютные значения коэффициента корреляции Спирмена (ККС) превышали 0,7 (Рис. 3).

Мультиколлинеарность зарегистрировали у 3 предикторов: систолическое АД, диастолическое АД и среднее АД (ККС>0,7). Среднее и диастолическое АД (2 из 3 переменных) исключили из последующего моделирования. Выбор систолического АД являлся эмпирическим, так как это наиболее важный параметр, который отражает сердечную функцию и сопротивление стенок кровеносных сосудов.

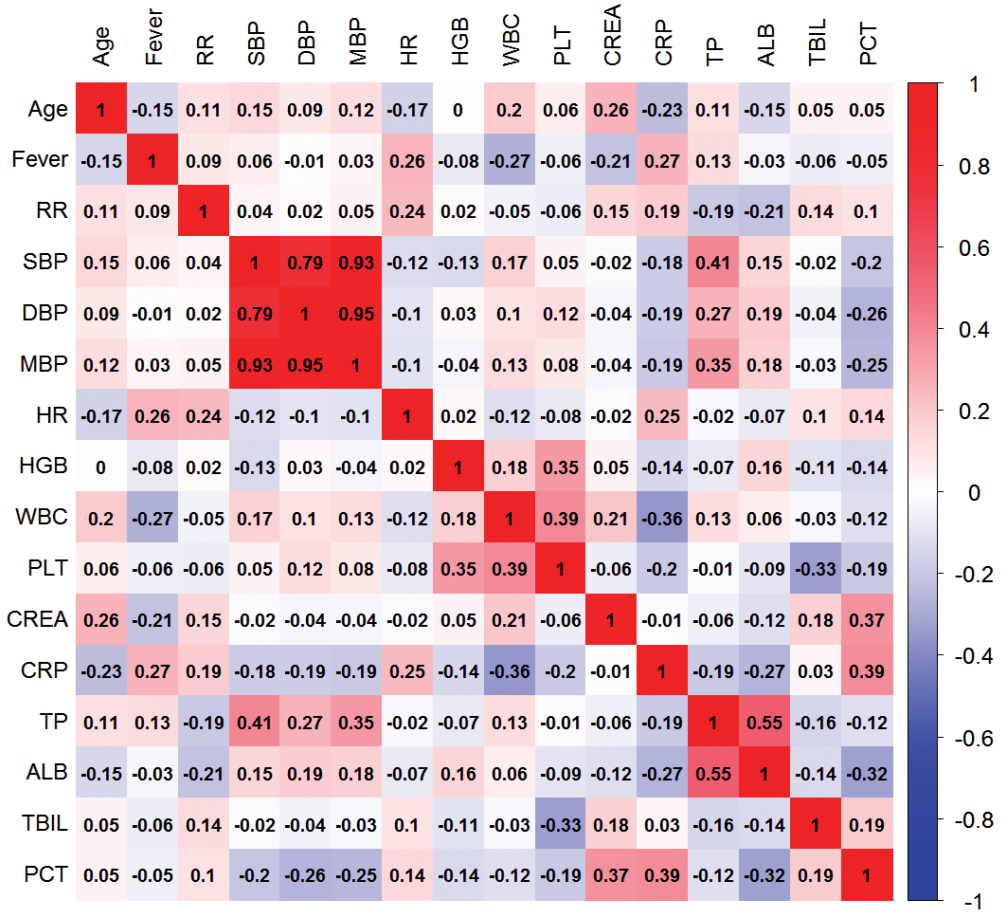

Рисунок 3 — Тепловая корреляционная матрица предикторов. *Age — возраст, Fever — повышение температуры, RR — частота дыхания, SBP — систолическое АД, DBP — диастолическое АД, MBP — среднее АД, HR — ЧСС, HGB — гемоглобин в крови, WBC — лейкоциты в крови, PLT — тромбоциты в крови, CREA — креатинин крови, CRP — С-реактивный белок, TP — общий белок крови, ALB — альбумин крови, TBIL — общий билирубин крови, PCT — прокальцитонин крови.

Критерий χ^2

Непараметрический критерий χ^2 применили для сравнения частот категориальных предикторов в группах с разным исходом госпитализации в ОИТ (табл. 2). Статистически значимые отличия по частоте летальности между предикторами использовали в качестве критериев их селекции.

Однофакторный дисперсионный анализ и критерий Манна-Уитни

Дисперсионный анализ (Analysis of Variance или ANOVA) позволяет найти различия между выборочными средними для разных совокупностей [18]. Данный метод применили для оценки количественных предикторов (табл. 3).

Таблица 2 — Сравнение категориальных предикторов в группах выживших и умерших

Группа	Умершие	Выжившие	Хи²; p
Мужской пол:			
Да	37	75	4,2; 1
Нет	30	60	
Инотропная поддержка катехоламинами:			
Да	14	7	**10,2; 0,001**
Нет	53	128	
Гипоксемия:			
Да	39	46	**9,7; 0,001**
Нет	28	89	
Уровень сознания по шкале Глазго <15 баллов:			
Да	21	5	**28; <0,001**
Нет	46	130	
Бактериемия:			
Да	13	27	2,05; 1
Нет	54	108	

Таблица 3 — Сравнение количественных предикторов в группах выживших и умерших

Предиктор	Умершие (M±□)	Выжившие (M±□)	p
Возраст, лет	52,5±15,3	53,3±15,1	0,715
Температура тела, C^0	37±1,1	37,1±1	0,588
ЧД, в мин.	22,5±4	20,8±3,5	0,319
Систолическое АД, мм рт. ст.	111,6±22,6	118,1±18,6	**0,032**
ЧСС, в мин.	100,2±17,8	96±17	0,114
Гемоглобин, г/л	85,7±12,5	81,6±15,8	0,079
Лейкоциты, 10^9/л*	0,8 [0,01–318,5]*	0,4 [0,01–703]	0,643
Тромбоциты, 10^9/л*	24 [1–339]	35 [2–807]	**0,006**
Креатинин, мкмоль/л	100,5±43,5	85,1±40,3	**0,018**
С-реактивный белок, г/л	0,18±0,1	0,14±0,1	**0,026**
Общий белок, г/л	57±8,5	61,2±9,3	**0,002**
Альбумин, г/л	30,7±4,7	33,7±5,8	**<0,001**
Общий билирубин, мкмоль/л	16,5±8,8	13,1±7,6	**0,008**
Прокальцитонин, нг/мл*	1,69 [0,02–125,5]	0,6 [0,04–86,2]	**0,005**

Примечание: * — значения переменных представлены в виде медианы [минимум — максимум], сравнение по методу Манна-Уитни.

Предварительно данные очистили от выбросов для приближения их распределения к нормальному. Переменные, которые не соответствовали условиям близкого к нормальному распределения, сравнивали при помощи рангового критерия Манна-Уитни. Статистически значимые отличия между предикторами использовали в качестве условий для их селекции.

Таким образом, применение базовых статистических контролируемых методов позволило уменьшить общее число предикторов с 21 до 11.

Таблица 4 — Оценка предикторов в однофакторном регрессионном анализе по методу Хосмера-Лемешоу

Предиктор	Отношение шансов (95% ДИ)	p
Мужской пол	0,98 (0,54–1,77)	0,964
Возраст, лет	0,99 (0,97–1,01)	0,713
Температура тела, С0	0,92 (0,7–1,21)	0,586
ЧД, в мин.	1,06 (1,01–1,12)	**0,019**
Систолическое АД, мм рт. ст.	0,98 (0,96–0,99)	**0,018**
ЧСС, мин$^{-1}$	1,01 (1,002–1,03)	**0,022**
Поддержка катехоламинами (Да/Нет)	4,83 (1,84–12,64)	**0,001**
Гипоксемия (Да/Нет)	2,69 (1,47–4,92)	**0,001**
Уровень сознания по шкале Глазго <15 баллов	11,86 (4,23–33,3)	**<0,001**
Гемоглобин, г/л	1 (0,99–1,02)	0,319
Лейкоциты, 10^9/л	0,99 (0,99–1,003)	0,915
Тромбоциты, 10^9/л	0,99 (0,99–1)	**0,085**
Креатинин, мкмоль/л	1 (0,99–1)	**0,068**
С-реактивный белок, г/л	9 (0,93–87,08)	**0,057**
Общий белок, г/л	0,95 (0,92–0,98)	**0,004**
Альбумин, г/л	0,92 (0,87–0,97)	**0,003**
Общий билирубин, мкмоль/л	1,01 (1,005–1,03)	**0,006**
Прокальцитонин, нг/мл	1,01 (0,99–1,03)	**0,067**
Бактериемия (Да/Нет)	0,96 (0,46–2,01)	0,92

Однофакторный регрессионный анализ

Однофакторный регрессионный анализ выполнили поочередно со всеми предикторами за исключением коллинеарных (диастолическое и среднее АД), пороговое значение альфа — 0,25 (табл. 4).

Рекурсивное удаление признаков (RFE)

Согласно методу RFE, наиболее значимыми предикторами в прогнозе летальности оказались: уровень сознания по шкале Глазго <15 баллов, концентрация общего белка крови, ЧД, инотропная поддержка катехоламинами и количество тромбоцитов в крови. Распределение всех переменных по степени важности представлено на рис. 4. Относительная степень важности представляет собой условный расчетный коэффициент, характеризующий вклад предиктора в прогноз целевой переменной. Предикторы, имеющие нулевую важность, могут быть исключены из модели без потери ее качества.

Регрессия «LASSO»

Цель регрессии «LASSO» состоит в том, чтобы получить подмножество предикторов, которое минимизирует ошибку предсказания для переменной отклика. «LASSO» делает это, накладывая ограничение на параметры модели через штрафную функцию лямбда, которая уменьшает дисперсию коэффициентов регрессии для некоторых переменных до нуля. Оптимальную величину параметра регуляризации лямбда, при которой ошибка прогноза минимальна, вычислили в результате 10-кратной кросс-валидации. СП по результатам регрессии «LASSO» позволила снизить общее число предикторов с 21 до 7: ЧСС, гипоксемия, количество тромбоцитов, концентрации общего белка, альбумина и общего билирубина в крови, уровень сознания по шкале Глазго в виде бинарной переменной, где 1 — количество баллов менее 15 и 0 — равно 15.

Алгоритм «Boruta»

Согласно алгоритму «Boruta» при p-уровне значимости ≤0,01 наиболее важными предикторами в прогнозе летальности оказались: уровень сознания по шкале Глазго, концентрация

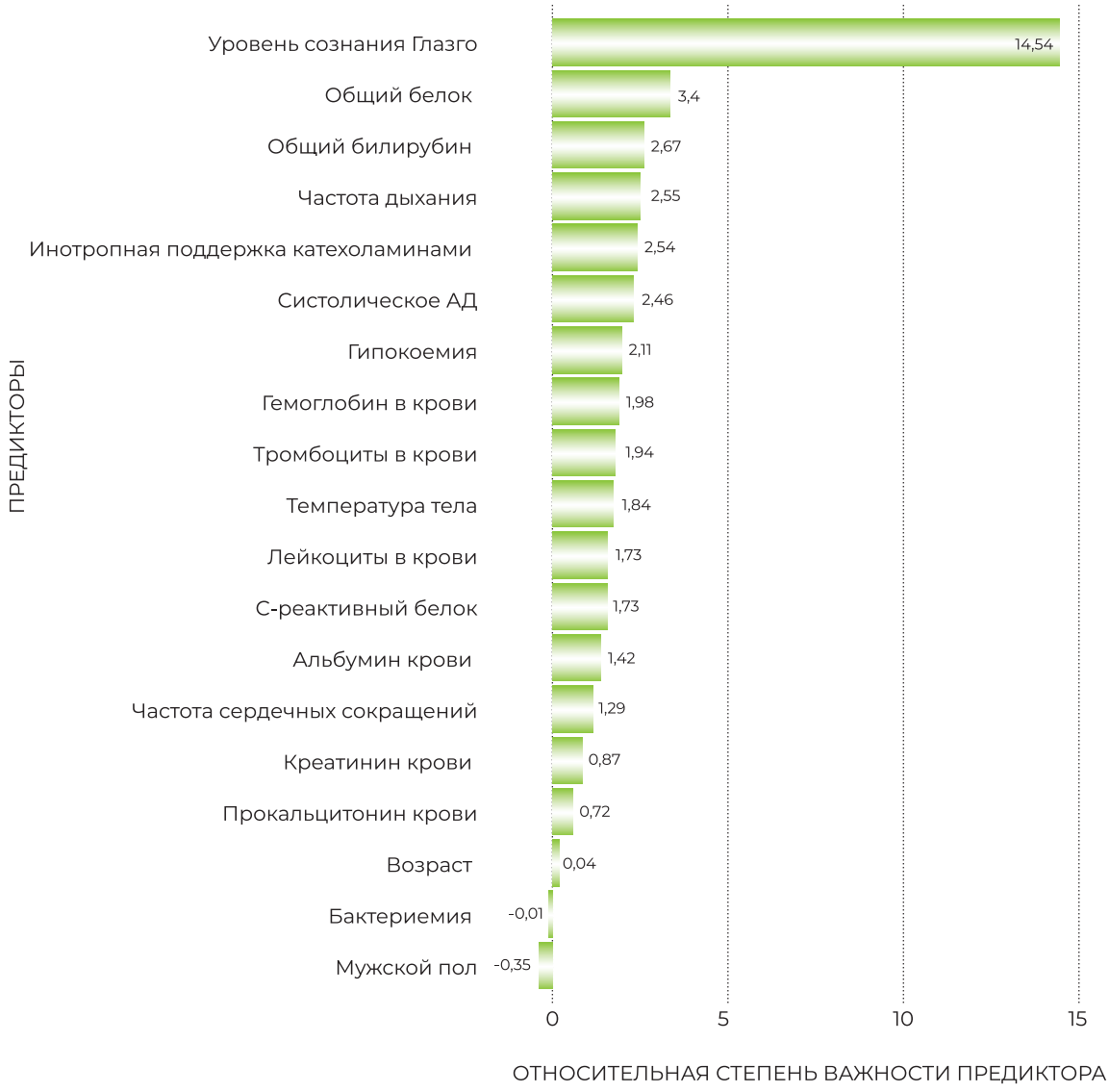

Рисунок 4 — Результаты селекции предикторов методом RFE.

общего белка крови, ЧД, инотропная поддержка катехоламинами, гипоксемия и систолическое АД (Рис. 5).

Пошаговая логистическая регрессия

Пошаговая регрессия — это пошаговое итеративное построение регрессионной модели, которое включает выбор независимых переменных для использования в окончательной модели. Метод заключается в последовательном добавлении или удалении потенциальных объясняющих переменных и проверку статистической значимости после каждой итерации. Начальный список состоял из 19 предикторов, 6 из которых

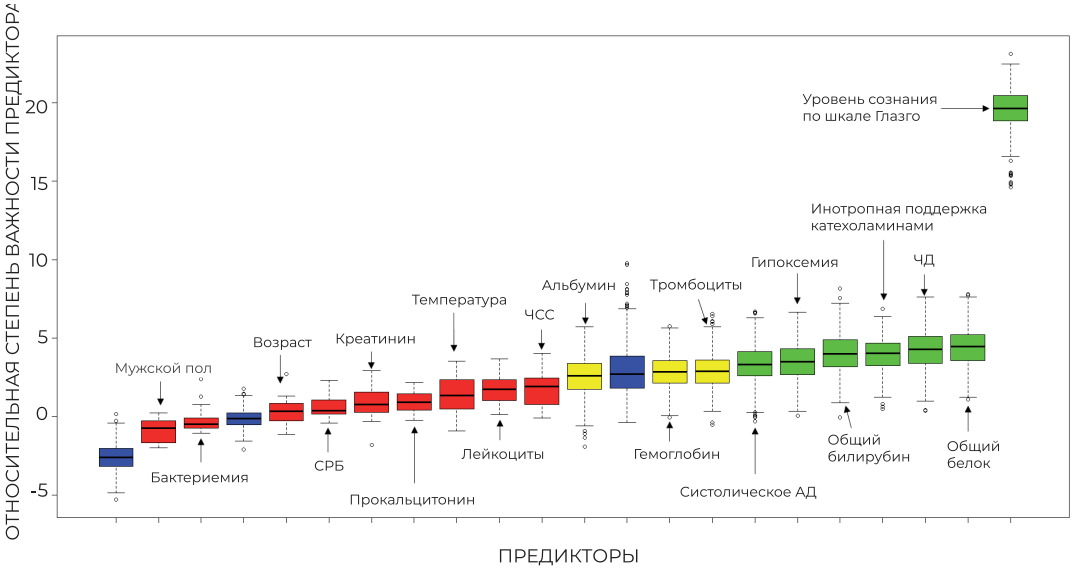

Рисунок 5 — Результаты селекции предикторов методом «Boruta».
*** — зеленым цветом выделены предикторы, имеющие наибольшую важность, красным — не значимые переменные, желтым — факторы с промежуточным значением, синим — «теневые предикторы».**

Таблица 5 — Сводные данные по методам селекции предикторов

Предиктор	Базовая статистика	Однофакторный регрессионный анализ	RFE	Регрессия «LASSO»	Метод «Boruta»	Пошаговая регрессия
Мужской пол						
Возраст, лет						
Температура тела, C^0						
ЧД, $мин^{-1}$		+	+		+	
Систолическое АД, мм рт. ст.	+	+			+	
ЧСС, $мин^{-1}$		+		+		+
Инотропная поддержка катехоламинами (Да/Нет)	+	+	+		+	
Гипоксемия (Да/Нет)	+	+		+	+	+
Уровень сознания по шкале Глазго <15 баллов	+	+	+	+	+	+
Гемоглобин, г/л						
Лейкоциты, 10^9/л						
Тромбоциты, 10^9/л	+	+	+	+		+
Креатинин, мкмоль/л	+	+				
С-реактивный белок, г/л	+	+				
Общий белок, г/л	+	+	+	+	+	+
Альбумин, г/л	+	+		+		
Общий билирубин, мкмоль/л	+	+		+	+	+
Прокальцитонин, нг/мл	+	+				
Бактериемия (Да/Нет)						

Примечание: * — знаком «+» обозначены выбранные предикторы.

Таблица 6 — Сравнение точности моделей логистической регрессии с разными наборами предикторов

Метод селекции предикторов	Конечное число предикторов	AIC
Базовая статистика (Хи^2, ANOVA, критерий Манна-Уитни)	11	219,2
Однофакторный регрессионный анализ	13	219,2
Рекурсивное удаление признаков (RFE)	5	217,9
Регрессия «LASSO»	7	208,6
Метод «Boruta»	7	208,6
Пошаговая регрессия	6	208,6

были включены в финальную модель логистической регрессии: ЧСС, гипоксемия, количество тромбоцитов, общий белок, общий билирубин и уровень сознания по шкале Глазго. В табл. 5 представлены сводные данные по СП, полученные с применением разных методов.

Моделирование

Задачей заключительного этапа исследования являлось построение моделей бинарной логистической регрессии с использованием различных наборов предикторов и сравнения точности прогнозирования. Качество моделей оценивали по значению критерия AIC (табл. 6).

Наборы предикторов, полученные при помощи метода «LASSO» случайного леса, а также методом пошаговой регрессии, позволили построить наиболее точные прогностические модели (минимальное значение AIC). Базовые методы статистического анализа и однофакторный регрессионный анализ по методу Хосмера-Лемешоу в СП оказались наименее эффективными.

ОБСУЖДЕНИЕ

Выбор признаков — важный этап решения задач прогнозирования и классификации, что подтверждается проведенным научным экспериментом. СП наиболее актуальна при работе с большими данными и сотнями переменных, она также может быть полезна при анализе данных небольшого размера. Не существует одного наилучшего метода СП, каждый из них должен сопоставляться с данными, типом модели и задачами научного исследования. Оптимальный способ — использовать систематические контролируемые эксперименты, чтобы выяснить, какие методы СП в сочетании с выбранной исследователем моделью обеспечивают наилучшую производительность в конкретном случае.

Сложные методы СП, такие как RFE, регрессия «LASSO» и «Boruta» наиболее полезны при большом количестве потенциальных предикторов. Проведенное нами исследование подтвердило их эффективность на небольшой по размерам выборке (табл. 6). Альтернативный вариант –пошаговая регрессия, которая выбирает наилучшую модель в процессе комбинации предикторов между собой, хотя и не перебирает все возможные варианты сочетания независимых переменных.

Таким образом, СП — универсальный этап прогнозной аналитики, который не зависит от методов моделирования, будь то логистическая регрессия или любой другой. В то же время СП — как правило, многократно повторяемый эксперимент с целью поиска оптимального решения. Применение методов СП не только помогает выбрать предикторы для последующего моделирования, но часто существенно сокращает их количество, отсеивая неинформативные.

Выбор метода СП должен базироваться на точно сформулированной научной задаче, типе данных и знаниях исследователя в области методов статистического анализа. Далеко не всегда требуется использование сложных способов СП, но их полное игнорирование может существенно снизить качество полученных результатов.

Список сокращений

АД — артериальное давление
ККС — коэффициент корреляции Спирмена
ОИТ — отделение интенсивной терапии
СП — селекция предикторов
ЧД — частота дыхания
ЧСС — частота сердечных сокращений
Age — возраст
AIC — информационный критерий Акаике
ALB — альбумин крови
ANOVA — дисперсионный анализ
CREA — креатинин крови
CRP — С-реактивный белок
DBP — диастолическое артериальное давление
Fever — повышение температуры
LASSO — оператор наименьшей абсолютной усадки и выбора
MBP — среднее артериальное давление
HR — частота сердечных сокращений
HGB — гемоглобин в крови
PCT — прокальцитонин
PLT — тромбоциты
RFE — рекурсивное удаление признаков
RR — частота дыхания
SBP — систолическое артериальное давление
TBIL — общий билирубин крови
TP — общий белок крови
WBC — лейкоциты в крови

ЛИТЕРАТУРА/REFERENCES

1. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ (Clinical research ed.). 2009; 338: b605. doi: 10.1136/bmj.b605.
2. van Beek PE, Andriessen P, Onland W, Schuit E. Prognostic Models Predicting Mortality in Preterm Infants: Systematic Review and Meta-analysis. Pediatrics. 2021; 147(5): e2020020461. doi: 10.1542/peds.2020-020461.
3. Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? Journal of the American Medical Informatics Association: JAMIA. 2019; 26(12): 1651-1654. doi: 10.1093/jamia/ocz130.
4. Andrade C. Sample Size and its Importance in Research. Indian Journal of Psychological Medicine. 2020; 42(1): 102-103. doi: 10.4103/IJPSYM.IJPSYM_504_19.
5. Pourhoseingholi MA, Vahedi M, Rahimzadeh M. Sample size calculation in medical studies. Gastroenterology and Hepatology from Bed to Bench. 2013; 6(1): 14-17.
6. Chowdhury MZI, Turin TC. Variable selection strategies and its importance in clinical prediction modelling. Family Medicine and Community Health. 2020; 8(1): e000262. doi: 10.1136/fmch-2019-000262.
7. Chen R-C, Dewi C, Huang S-W, Caraka RE. Selecting critical features for data classification based on machine learning methods. Journal of Big Data. 2020; 7(1): 52. doi: 10.1186/s40537-020-00327-4.
8. Staartjes VE, Kernbach JM, Stumpo V, van Niftrik CHB, Serra C, Regli L. Foundations of Feature Selection in Clinical Prediction Modeling. Acta Neurochirurgica. Supplement. 2022; 134: 51-57. doi: 10.1007/978-3-030-85292-4_7.
9. Li L. Dimension reduction for high-dimensional data. Methods in Molecular Biology (Clifton, N.J.). 2010; 620: 417-434. doi: 10.1007/978-1-60761-580-4_14.
10. Ameringer S, Serlin RC, Ward S. Simpson's Paradox and Experimental Research. Nursing research. 2009; 58(2): 123-127. doi: 10.1097/NNR.0b013e318199b517.
11. Kim JH. Multicollinearity and misleading statistical results. Korean Journal of Anesthesiology. 2019; 72(6): 558-569. doi: 10.4097/kja.19087.
12. Zhang Z. Variable selection with stepwise and best subset approaches. Annals of Translational Medicine. 2016; 4(7): 136. doi: 10.21037/atm.2016.03.35.
13. Tibshirani R. The lasso method for variable selection in the Cox model. Statistics in Medicine. 1997; 16(4): 385–395. doi: 10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3.

14. Rigatti SJ. Random Forest. Journal of Insurance Medicine (New York, N.Y.). 2017; 47(1): 31-39. doi: 10.17849/insm-47-01-31-39.1.
15. Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. Briefings in Bioinformatics. 2019; 20(2): 492-503. doi: 10.1093/bib/bbx124.
16. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. Journal of Clinical Epidemiology. 1996; 49(8): 907-916. doi: 10.1016/0895-4356(96)00025-x.
17. Dziak JJ, Coffman DL, Lanza ST, Li R, Jermiin LS. Sensitivity and specificity of information criteria. Briefings in Bioinformatics. 2020; 21(2): 553-565. doi: 10.1093/bib/bbz016.
18. Thompson HW, Mera R, Prasad C. The Analysis of Variance (ANOVA). Nutritional Neuroscience. 1999; 2(1): 43-55. doi: 10.1080/1028415X.1999.11747262.