УДК: 616.24-006:036.65+004.8 DOI: 10.52532/2521-6414-2022-3-65-4-11

ПРИМЕНЕНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РАЗРАБОТКИ МОДЕЛИ ПРОГНОЗИРОВАНИЯ РЕЗУЛЬТАТОВ ВЫЖИВАЕМОСТИ БОЛЬНЫХ РАКОМ ЛЕГКИХ В РК

В.А МАКАРОВ[1,2], Д.Р. КАЙДАРОВА[3], С.Е. ЕСЕНТАЕВА[4], Ж. КАЛМАТАЕВА[2], М.Е. МАНСУРОВА[2], Н. КАДЫРБЕК[2], Р.Е. КАДЫРБАЕВА[2], С.Т. ОЛЖАЕВ[1], И.И. НОВИКОВ[1]

[1]КГП на ПХВ «Алматинская Региональная Многопрофильная Клиника», Алматы, Республика Казахстан;
[2]НАО «Казахский Национальный Университет им. аль-Фараби», Алматы, Республика Казахстан;
[3]АО «Казахский Научно-Исследовательский Институт Онкологии и Радиологии», Алматы, Республика Казахстан;
[4]НУО УО «Казахстанско-Российский медицинский университет», Алматы, Республика Казахстан

АННОТАЦИЯ

Актуальность: *В ряде исследований было показано, что модели, созданные с помощью искусственного интеллекта, являются более точными, чем обычная система стадирования TNM, поскольку они строятся на анализе большого объема данных, отражающих как биологические, так и клинические особенности течения болезни. На этом основании модели, созданные с помощью машинного обучения, были рекомендованы в качестве альтернативных или дополняющих TNM классификацию прогностических инструментов.*

Цель исследования *– оценить прогностическую значимость ряда клинико-морфологических факторов и применить алгоритмы машинного обучения для прогнозирования результатов общей выживаемости больных с раком легких.*

Методы: *Проведен анализ истории болезни пациентов с раком легкого (n=19379) из базы данных ЭРОБ за 2014-2018 гг., произведена оценка влияния факторов риска на общую выживаемость по методу Каплана-Мейера. Примененные в работе алгоритмы машинного обучения (Random Forest Classifier, Gradient Boosting Classifier, Logistic Regression Model, Decision Tree Classifier, K Nearest Neighbors (KNN) Classifier) реализованы на языке программирования Python.*

Результаты: *В нашем исследовании были проанализированы истории болезни 19 379 пациентов. На момент исследования среди мужчин были живы 6 171 больных (39,8%), при этом медиана выживаемости составила 8,3 месяцев (SE – 0,154 месяцев, 95% ДИ – 7,96-8,56). Среди женщин были живы 1 962 больных (49,5%), при этом медиана выживаемости составила 15,43 месяцев (SE – 1,0 месяц, 95% ДИ – 13,497-17,363). У большинства (61,4%) пациентов НМРЛ был диагностирован в распространенной стадии: у 9 189 человек (47,4%) – на III стадии, у 4 655 (24%) – на IV стадии. Оценка достоверности различий в медиане выживаемости ($\chi 2=3991,6$, $p=0,00$) указывает на прогностическую значимость и влияние стадии опухолевого процесса на выживаемость больных.*

Заключение: *Модели машинного обучения позволяют прогнозировать риск развития летального исхода больных как после хирургического лечения, так и после постановки на учет в базу данных ЭРОБ. Создание пациент-ориентированных систем поддержки принятия врачебных решений позволяет выбрать оптимальные стратегии адъювантной терапии, диспансерного наблюдения и частоты диагностических исследований.*

Ключевые слова: *рак легкого, прогностическая значимость, машинное обучение, рецидивы, общая выживаемость.*

Введение: На протяжении последних десятилетий онкопатология органов грудной клетки является главной причиной онкологических заболеваний и причиной большинства случаев смерти. Однако, несмотря на выявление болезни на ранних стадиях, часть пациентов все-таки умирает от рецидива заболевания. По данным R. Maeda, частота рецидивов у радикально пролеченных пациентов составляет почти 10% [1]. Определение рисков рецидивов и/или смертельных исходов у больных НМРЛ остается важной малоизученной проблемой. Современная система стадирования опухолей (7-ая и 8-ая классификация TNM) является наиболее часто используемым инструментом прогнозирования для НМРЛ. Тем не менее, данная классификация отражает не все важные клинические и патологические предикторы, поэтому бывает бессильна для определения персонализированного подхода в прецизионной медицине [2-4]. В ряде исследований было показано, что модели, созданные с помощью искусственного интеллекта, являются более точными, чем обычная система стадирования TNM, поскольку они строятся на анализе большого объема данных, отражающих как биологические, так и клинические особенности течения болезни [5]. На этом основании модели, созданные с помощью машинного обучения, были рекомендованы в качестве альтернативных или дополняющих TNM классификацию прогностических инструментов [6]. Обзор литературы указывает на успешное применение следующих алгоритмов машинного обучения: *Random Forest Classifier*, *Gradient Boosting Classifier, Logistic Regression Model, Decision Tree Classifier, K Nearest Neighbors (KNN) Classifier* в классификации больных РЛ по группам риска [7-12] и прогнозировании выживаемости больных РЛ [13-14].

Цель исследования – оценить прогностическую значимость ряда клинико-морфологических факторов и применить алгоритмы машинного обучения для прогнозирования результатов общей выживаемости больных с раком легких.

Материалы и методы: Проведен анализ форм 030-6/у С34 – рак легкого (n=19379) из базы данных ЭРОБ за 2014-2018 гг. Произведена оценка влияния факторов риска (пол, возраст, TNM, гистология, локализация метастатических очагов) на общую выживаемость по методу Каплана-Мейера. Создание базы данных осуществлялось в программе Microsoft Excel. Соответственно обучающий набор данных для построения моделей прогнозирования включает в себя 19379 наблюдений и 15 факторов. Нами определено три группы риска: Группа 1 – выживаемость от 0 до 12 мес., Группа 2 – 12-24 мес., Группа 3 – 24-72 мес., соответственно.

Примененные в работе алгоритмы машинного обучения *(Random Forest Classifier, Gradient Boosting Classifier, Logistic Regression Model, Decision Tree Classifier, K Nearest Neighbors (KNN) Classifier)* реализованы на языке программирования *Python*. Результаты оценивались с помощью построения матрицы ошибок, расчета метрик классификации: доли правильно классифицированных объектов *(accuracy)* при обучении и проверке *(validation)*, точности измерений *(precision)*, полноты *(recall)* Каппа-Коэна.

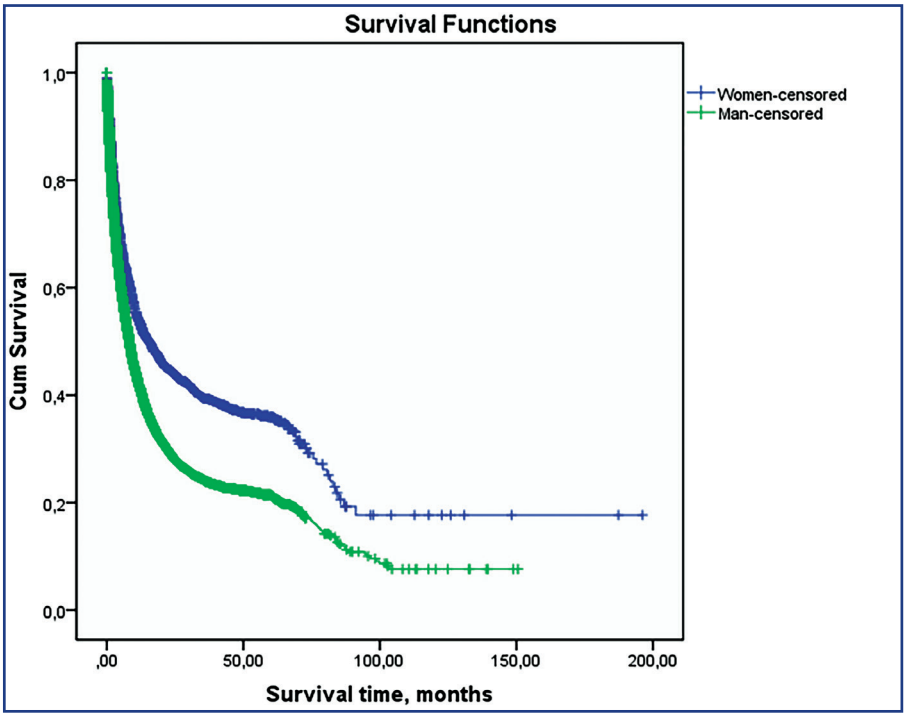
Результаты:

Оценка влияния фактора пола на выживаемость больных РЛ в РК

В нашем исследовании было проанализировано 19 379 больных, в том числе 15 494 мужчин (79,95%) и 3 885 женщин (20,04%).

На момент исследования среди мужчин были живы 6 171 больных (39,8%), при этом медиана выживаемости составила 8,3 месяцев (SE – 0,154 месяцев, 95% ДИ – 7,96-8,56). Одногодичная выживаемость среди мужчин составляла 44% (SE – 0,44), 2-летняя – 31% (SE – 4,4), 3-летняя – 26% (SE – 0,47), 4-летняя – 24% (SE – 0,49), при этом 5-летняя выживаемость достигала 23% (SE – 0,51).

Среди женщин были живы 1 962 больных (49,5%), при этом медиана выживаемости составила 15,43 месяцев (SE – 1,0 месяц, 95% ДИ – 13,497-17,363). Одногодичная выживаемость среди женщин равнялась 55% (SE – 0,84), 2-летняя – 45% (SE – 0,9), 3-летняя – 40% (SE – 0,95), 4-летняя – 38% (SE – 1,0). 5-летняя выживаемость составляла 37% (SE – 1,03) (Рисунок 1).



Легенда: ось Y - Кумулятивные показатели выживаемости; ось X - Срок выживания, месяцы

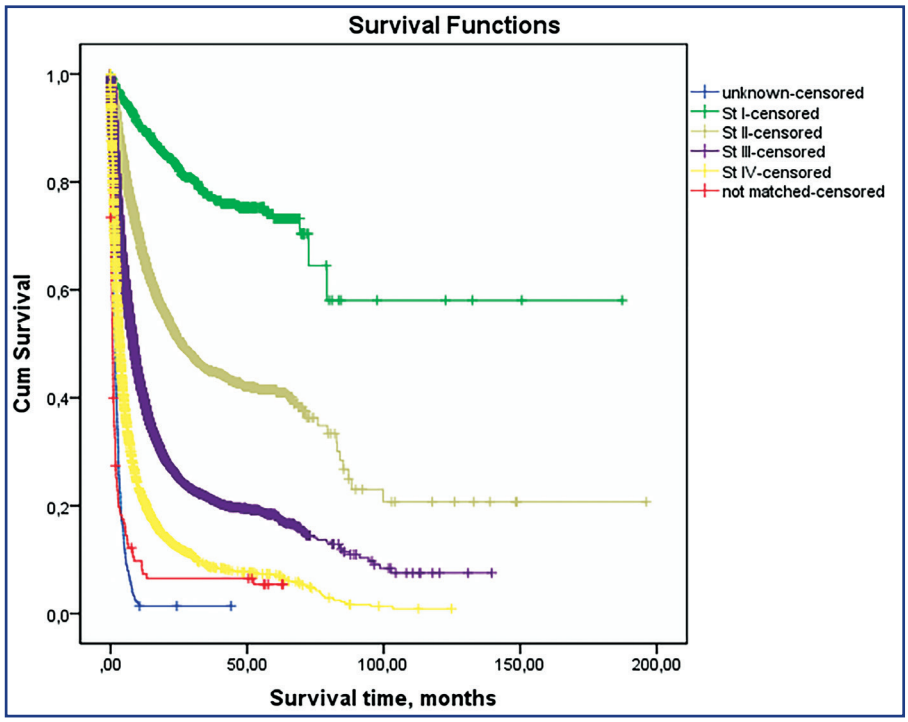
Рисунок 1 – Общая выживаемость больных в зависимости от пола, по методу Каплана-Мейера

Таким образом, очевидно, что на нашем наборе данных мужской пол является фактором риска по выживаемости при РЛ. Отмечается существенная статистически достоверная разница в медиане выживаемости мужчин и женщин с РЛ: $\chi 2=219,03$, p=0,00.

Влияние стадии опухолевого процесса на отдаленные результаты больных РЛ

У большинства (61,4%) пациентов НМРЛ был диагностирован в распространенной стадии: у 9 189 человек (47,4%) – на III стадии, у 4 655 (24%) – на IV стадии.

Среди больных с I стадией на конец 2018 г. были живы 845 больных (81,5%). При этом медиана не была достигнута: средние показатели выживаемости составили 125,6 месяцев, SE – 9,6 месяцев, 95% ДИ – 106,7-144,5. На момент завершения периода исследования 2 366 (57,1%) больных со II стадией были живы. Их медиана выживаемости соответствовала 26,1 месяцам, SE – 1,4 месяца, 95% ДИ – 23,3-28,8. Из 9 189 больных с III стадией в живых осталось 3 687 (40,1%) с медианой выживаемости 8,3 месяцев, SE – 0,2 месяца, 95% ДИ – 8,0-8,7. При этом к концу 2018 г. осталась в живых только четвертая часть пациентов с IV стадией заболевания – 1 183 (25,4%). Медиана выживаемости в этой группе составила 3,3 месяца, SE – 0,1 месяца, 95% ДИ – 3,1-3,5 (рисунок 2).



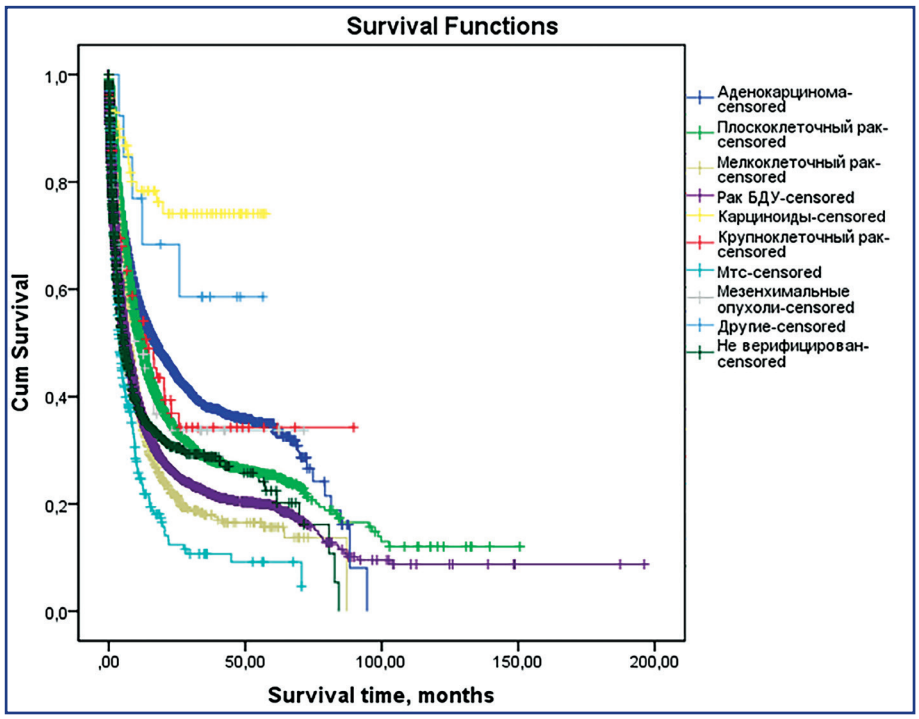
Легенда: ось Y - Кумулятивные показатели выживаемости; ось X - Срок выживания, месяцы

Рисунок 2 – Общая выживаемость больных в зависимости от стадии РЛ по методу Каплана-Мейера

Оценка достоверности различий в медиане выживаемости (χ2=3991,6, p=0,00) указывает на прогностическую значимость и влияние стадии опухолевого процесса на выживаемость больных.

Влияние морфологического типа опухоли на выживаемость больных раком легкого в РК

Из 19 379 больных, у которых был диагностирован РЛ в 2014-2018 гг., на долю аденокарциномы пришлось 18,5% (3 579 человек). На конец 2018 г. были живы 1 738 (48,6%) больных, при этом медиана выживаемости составила 17,1 месяцев, SE – 0,9 месяцев, 95% ДИ – 15,2-19,1. Среди больных плоскоклеточным раком, на долю которых пришлось 27,0% (5231), были живы 2 254 пациента (43,1%), при этом медиана выживаемости составила 11,6 месяцев, SE – 0,3 месяца, 95% ДИ – 10,9-12,3 (Рисунок 2).



Легенда: ось Y - Кумулятивные показатели выживаемости; ось X - Срок выживания, месяцы

Рисунок 3 – Общая выживаемость больных в зависимости от морфологического типа опухоли по методу Каплана-Мейера

Мелкоклеточный рак (МРЛ) был диагностирован в 1 091 (5,6%) случаях. На конец 2018 г. были живы 377 (34,6%) больных, при этом медиана выживаемости составила 7,2 месяцев, SE – 0,3 месяца, 95% ДИ – 6,5-7,99. Формы рака легкого без дополнительного уточнения (БДУ) были выявлены в 7 643 (39,4%) случаях, живы были 2 922 (38,2%) больных, при этом медиана выживаемости составила 6,2 месяцев, SE – 0,2 месяца, 95% ДИ – 5,7-6,6. Среди пациентов с аденокарциномой легких, одно-, двух-, трех-, четырехгодичная выживаемость составили 57% SE1, 45% SE1, 39% SE1 и 37% SE1, соответственно. Показатели пятилетней выживаемости составили 36% SE1. У пациентов с плоскоклеточным раком легких показатели одно-, двух-, трех- и четырехлетней выживаемости были несколько ниже и составили 51% SE1, 35% SE1, 30% SE1 и 28% SE1, соответственно. Пятилетняя выживаемость была равна 27% SE1. Для МРЛ, основные показатели выживаемости оказались еще более низкими по сравнению с НМРЛ: одногодичная выживаемость составила 39% SE2, двухлетняя – 24% SE2, трехлетняя – 21% SE 2 и четырехлетняя выживаемость – 19% SE 2. Уровень пятилетней выживаемости не превысил 20%-ного порога и составил 18% SE2.

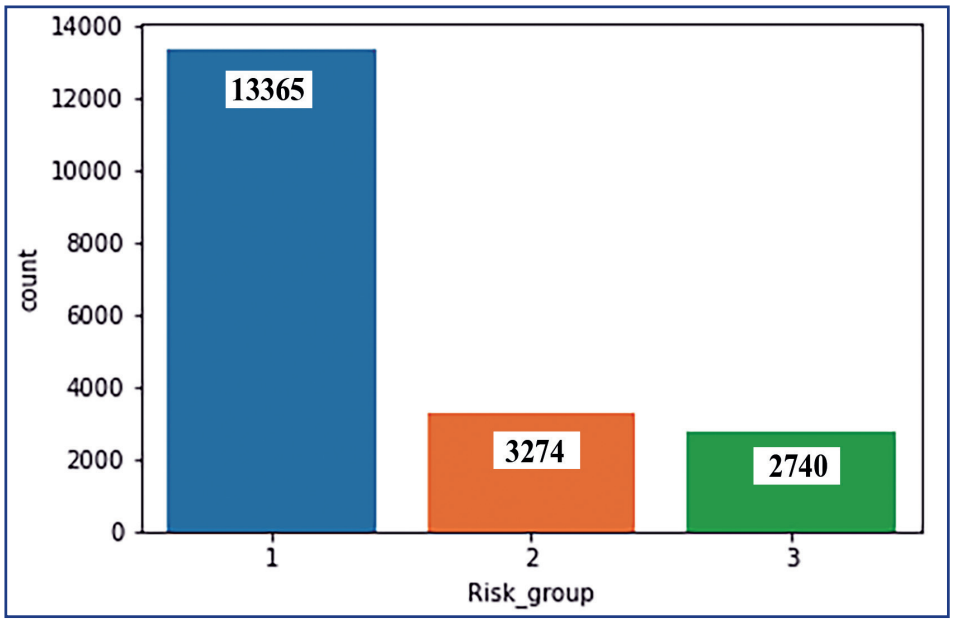
Уровень показателей выживаемости у больных с недифференцированным раком легких перекликался с таковым при МРЛ: 40% SE1, 28% SE1 и 24% SE1 для одно-, двух- и трёхлетней выживаемости, и 22% SE1 – для четырех- и пятилетней выживаемости. Одногодичная выживаемость среди больных с карциноидами составила 78% SE5. Двух-, трёх-, четырех- и пятилетняя выживаемость составила 74% SE6.

Таким образом, обязательная морфологическая идентификация злокачественных опухолей легких помогает не только в выборе тактики лечения и подбора адекватной противоопухолевой лекарственной терапии, но и способствует определению прогноза заболевания. Выявленная существенная разница в медиане выживаемости среди пациентов с различными морфологическими формами рака легкого позволяет говорить о прогностической значимости морфологического фактора (статистически разница между этими показателями оказалась достоверной, $\chi 2=623,4$ $p=0,000$).

Прогнозирование метки выживаемости больных РЛ из базы данных ЭРОБ с помощью «модели машинного обучения».

После проведения оценки потенциально значимых предикторов из базы данных ЭРОБ была сформирована обучающая выборка. Модели, созданные с помощью «модели машинного обучения», автоматически классифицируют больных с учетом многофакторных данных.

Как видно из рисунка 4, наибольшее количество больных наблюдалось в 1-ой группе риска.



Легенда: ось Y - количество пациентов; ось X - Группа риска

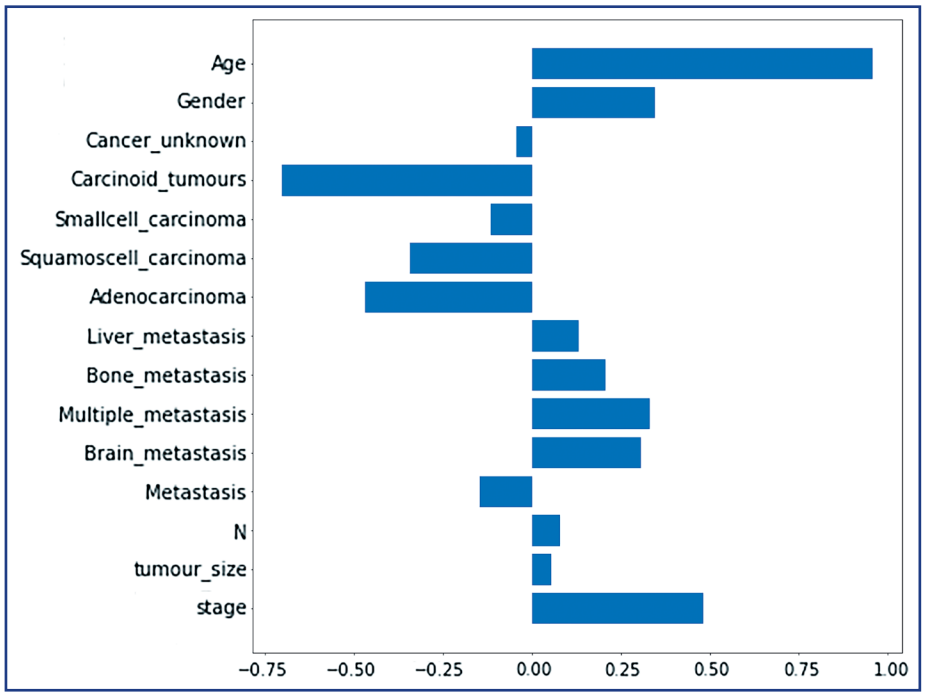
Рисунок 4 – Распределение больных РЛ по группам риска

При построении «модели машинного обучения» для нас было не ясно, какие из параметров действительно важны для нее, а какие являются избыточными (или шумовыми). Исключение избыточных параметров позволяет лучше интерпретировать данные, повысить точность модели. Во время обучения модели мы оптимизировали список отобранных признаков для повышения точности моделирования. Основными предикторами (~5%) в данных моделях были определены стадия и размер опухоли, уровень поражения лимфатических узлов (N), а также возраст пациентов (таблица 1).

Разработанные модели машинного обучения показали высокую долю правильно сгруппированных объектов классификации, т.е. высокую точность модели (*accuracy*). Наибольшая точность предсказаний на обучающем наборе была достигнута с помощью алгоритмов дерева решений (*Decision Tree*) (0,86), градиентного бустинга (*Gradient Boosting*) (0,72) и случайного леса (*Random Forest*) (0,70). При проверке (*validation*) полученных моделей показатели точности *(accuracy)* были следующими: для алгоритма градиентного бустинга – 0,70, случайного леса – 0,70, логистической регрессии – 0,69 (рисунок 5, таблица 2).

Таблица 1 – Расчет важности признаков в алгоритмах, %

Признак	Алгоритм		
	Decision Tree Classifier	Random Forest Classifier	Gradient Boosting Classifier
Stage	16,0	34,7	59,3
Tumour_size	8,6	17,2	6,0
N	9,1	17,4	8,2
Metastasis	3,7	12,9	4,3
Brain_metastasis	0,7	0,3	0,3
Multiple_metastasis	2,1	0,9	0,2
Bone_metastasis	1,0	0,2	0,3
Liver_metastasis	1,7	0,2	0,4
Adenocarcinoma	2,7	4,3	4,0
Squamoscell_carcinoma	3,4	1,6	1,4
Smallcell_carcinoma	2,0	0,2	0,1
Carcinoid_tumours	0,4	0,4	0,8
Cancer_unknown	2,5	3,4	2,6
Gender	3,6	2,9	4,3
Age	42,4	3,4	7,8
Итого	**100**	**100**	**100**



Легенда: ось Y - фактор прогноза; ось X - показатели регрессии

Рисунок 5 – Модель прогнозирования выживаемости, созданная с помощью алгоритма логистической регрессии

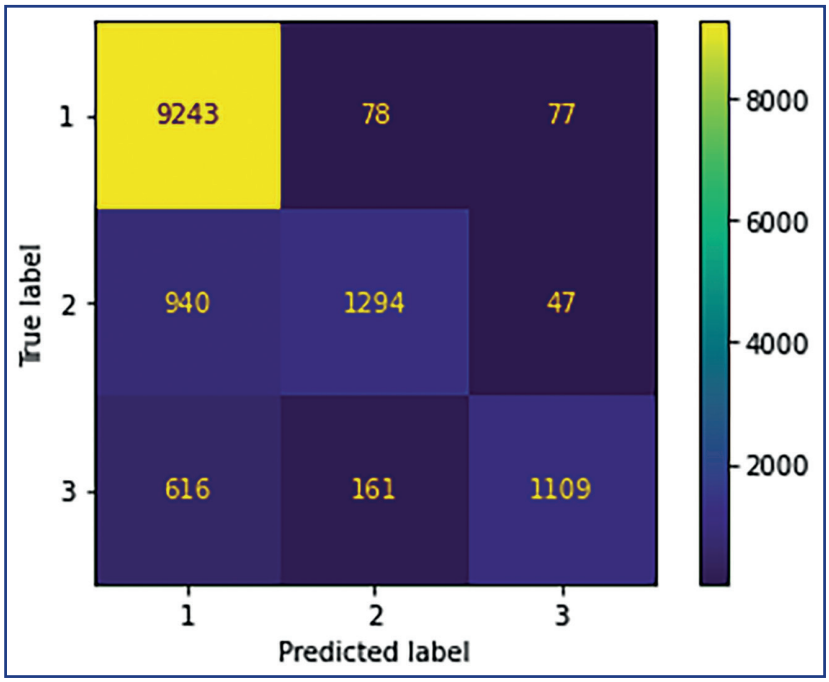
Таблица 2 – Показатели точности алгоритмов машинного обучения при обучении и проверке

Алгоритмы машинного обучения	Точность при обучении	Точность при тестировании
DecisionTreeClassifier	0,86	0,63
RandomForestClassifier	0,71	0,70
GradientBoostingClassifier	0,72	0,70
LogisticRegressionModel	0,70	0,69
K NearestNeighborsClassifier	0,75	0,68

Алгоритм дерева решений на данном наборе данных показал наилучшие характеристики (точность (*accuracy)* при обучении – 0,86, при проверке – 0,63). После подбора оптимальных параметров для модели, а именно {'C': 100, 'penalty': 'l2', 'solver': 'liblinear'}, точность (*accuracy)* при проверке составила 69%. Качество данной модели было проверено с помощью матрицы ошибок (рисунок 6), по данным которой значение степени точности измерения (*precision)* составило 0,71, полноты (*recall*) – 0,87. Показатель меры согласованности – k-Коэна – составил 0,66, что говорит о хорошем потенциале данного подхода. Другие значения составили: истинно положительная пропорция (true positive rate, TPR) – 0,98, ложно положительная пропорция (false positive rate, FPR) – 0,06, специфичность – 0,94, площадь под кривой (area under curve, AUC) – 0,98.

Прогнозирование метки выживаемости для больных, состоящих в базе данных ЭРОБ (за период 2014-2018 гг., 19 379 больных и 15 факторов) с помощью машинного

обучения позволяет сделать следующее заключение. Наилучшие модели были созданы с помощью алгоритмов машинного обучения как случайный лес, градиентный бустинг, дерево решений, логистическая регрессия. При этом были достигнуты показатели точности (*accuracy)* 72% при обучении и 70% при проверке на тестовом наборе. Для Decision Tree Classifier мера точности *(accuracy*) в обучающем наборе составила 87%, тогда как проверка на тестовом наборе показала точность в 63%. После подбора оптимальных параметров для модели с использованием {'C': 100, 'penalty': 'l2', 'solver': 'liblinear'} точность (*accuracy)* при проверке составила 69%.



Легенда: ось X – Предсказанные значения; ось У – Реальные значения

Рисунок 6 – Матрица ошибок модели прогнозирования для больных из трех групп риска базы данных ЭРОБ, созданной с помощью алгоритма дерево решений

По данным матрицы ошибок степень точности измерения (*precision)* составила 0,71, полнота (*recall*) – 0,87, критерий Коэна – 0,66, TPR – 0,98, FPR – 0,06, специфичность – 0,94, AUC – 0,98.

В данном случае построенные модели достигли допустимых для применения показателей, в связи с чем данные модели признаны работающими.

Обсуждение: Система стадирования TNM, оставаясь важным и доказанным предиктором выживания, не охватывает всех параметров болезни, что ограничивает ее возможности, как прогностического фактора риска неблагоприятного исхода. В настоящем исследовании с помощью оценки влияния на выживаемость по методу Каплана-Мейера были изучены прогностические и предиктивные биомаркеры, влияющие на прогноз заболевания. Так, например, мужской пол определен как фактор риска, поэтому стоит отметить интерес в изучения данного фактора. Предположительно, ни сам пол, как генетический фактор риска, а скорее всего связь пола с образом жизни, поведенческими факторами, работой, отношением к здоровью и др.

Применение разработанных алгоритмов также позволяет идентифицировать подгруппы больных, нуждающихся в более интенсивном наблюдении и адъювантных режимах лечения. Стратификация риска по полученным данным может способствовать изменению установленных стандартов наблюдения и лечения в пользу выбора дальнейшей лекарственной терапии и интенсивности диспансерного наблюдения. Так, пациентам с высоким риском необходимо сокращать период диспансерного наблюдения, чтобы своевременно скорректировать методы лечения в соответствии с изменениями онкологического и функционального статуса. И всё же вопрос идентификации подгрупп пациентов с высоким риском рецидива и пользой от адъювантной терапии остаётся открытым, а отбор кандидатов на химиотерапию на основе единственного фактора риска может быть неэффективным, так как для полноценного прогнозирования необходимо учитывать все атрибуты заболевания и вес каждого фактора.

В данном исследовании «модели машинного обучения» показали оптимальное сочетание между прогнозом и фактическим наблюдением, что гарантирует воспроизводимость и надежность предложенной модели. Что еще более важно, предложенная модель соответствует когорте ЭРОБ.

Алгоритмы машинного обучения представляют собой более точную прогностическую модель по сравнению с системой стадирования TNM и разработанными ранее прогностическими моделями. Пользуясь данным инструментом, врачи смогут более точно спрогнозировать выживаемость отдельных пациентов после операции и определить подгруппы больных, которые нуждаются в конкретной стратегии лечения.

Выводы:

1. При анализе факторов, влияющих на выживаемость при РЛ, выявлено, что мужской пол является фактором риска (χ2=219,03, p=0,00), тогда как женский пол отнесен к факторам благоприятного прогноза.

2. Анализ клинико-морфологических факторов показал достоверное влияние на выживаемость при РЛ таких показателей, как стадия заболевания (χ2=3991,6, p=0,00) и морфологический тип опухоли (χ2=623,4 p=0,000).

3. Оценка достоверности различий в медиане выживаемости (χ2=3991,6, p=0,00) указывает на прогностическую значимость и влияние стадии РЛ на выживаемость.

4. Классификаторы Random Forest, Gradient Boosting, Decision Tree показали себя как приемлемые классификаторы в прогнозировании группы риска (метки) общей выживаемости при РЛ.

5. Оценка «модели машинного обучения» на тестовом наборе показала допустимые параметры для применения классификаторов Random Forest, Gradient Boosting, Decision Tree в качестве вспомогательного инструмента при принятии решений.

6. Для построения модели прогнозирования, помимо алгоритма, большое значение имеет качество данных. Данные должны быть точными и в большом количестве, и иметь нормальное (гауссовское) распределение по группам риска (классам).

Заключение: Модели машинного обучения позволяют прогнозировать риск развития летального исхода больных РЛ как после хирургического лечения, так и после постановки на учет в базу данных ЭРОБ. Создание пациент-ориентированных систем поддержки принятия врачебных решений позволяет выбрать оптимальные стратегии адъювантной терапии, диспансерного наблюдения и частоты диагностических исследований.

Список использованных источников:

1. Maeda R., Yoshida J., Ishii G., Aokage K., Hishida T., Nishimura M., Nishiwaki Y., Nagai K. Long-term outcome and late recurrence in patients with completely resected stage IA non-small cell lung cancer // J. Thorac. Oncol. – 2010. – Vol. 5. – P. 1246-1250. https://doi.org/10.1097/JTO.0b013e3181e2f247;

2. Amin M.B., Greene F., Edge S.B., Compton C.C., Gershenwald J.E., Brookland R.K., Meyer L., Gress D.M., Byrd D.R., Winchester D.P. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging // CA: Cancer J. Clin. – 2017. – Vol. 67(2). – P. 93-99. https://doi.org/10.3322/caac.21388;

3. Goldstraw P., Chansky K., Crowley J., Rami-Porta R., Asamura J., Eberhardt W.E.E., Nicholson A.G., Groome P., Mitchell A., Bolehack V., on behalf of the International Association for the Study of Lung Cancer Staging and Prognostic Factors Committee, Advisory Boards, and Participating Institutions. The IASLC lung cancer-staging project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer // J. Thorac. Oncol. – 2016. – Vol. 11(1). – P. 39-51. https://doi.org/10.1016/j.jtho.2015.09.009;

4. Rami-Porta R., Bolejack V., Crowley J., Ball D., Kim J., Lyons G., Rice T., Suzuki K., Thomas C.F. Jr., Travis W.D., Wu Y.-L., on behalf of the IASLC Staging and Prognostic Factors Committee, Advisory Boards, and Participating Institutions. The IASLC Lung Cancer Staging Project: Proposals for the Revisions of the T Descriptors in the Forthcoming Eighth Edition of the TNM Classification for Lung Cancer // J. Thorac. Oncol. – 2015. – Vol. 10(7). – P. 990-1003. https://doi.org/10.1097/JTO.0000000000000559;

5. Balachandran V.P., Gonen M., Smith J.J., DeMatteo R.P. Nomograms in oncology: more than meets the eye // Lancet Oncol. – 2015. – Vol. 16(4). – P. e173-180. https://doi.org/10.1016/S1470-2045(14)71116-7;

6. Kourou K., Exarchos K.P., Papaloukas C., Sakaloglou P., Exarchos T., Fotiadis D.I. Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis // Comput. Struct. Biotechnol. J. – 2021. – Vol. 19. – P. 5546-5555. https://doi.org/10.1016/j.csbj.2021.10.006;

7. Lynch C.M., Abdollahi B., Fuqua J.D., de Carlo A.R., Bartholomai J.A., Balgemann R.N., van Berkel V.H., Frieboes H.B. Prediction of lung cancer patient survival via supervised machine learning classification techniques // Int. J. Med. Inform. – 2017. – Vol. 108. – P. 1-8. https://doi.org/10.1016/j.ijmedinf.2017.09.013;

8. Ramroach S., Joshi A., John M. Optimisation of cancer classification by machine learning generates an enriched list of candidate drug targets and biomarkers // Mol. Omics. – 2020. – Vol. 16(2). – P. 113-125. https://doi.org/10.1039/c9mo00198k;

9. Levitsky A., Pernemalm M., Bernhardson B.M., Forshed J., Kölbeck K., Olin M., Henriksson R., Lehtiö J., Tishelman C., Eriksson L.E. Early symptoms and sensations as predictors of lung cancer: a machine learning multivariate model // Sci. Rep. – 2019. – Vol. 9(1). – Art. ID 16504. https://doi.org/10.1038/s41598-019-52915-x;

10. Zhang X., Wang J., Li J., Chen W., Liu C. CRlncRC: a machine learning-based method for cancer-related long noncoding RNA identification using integrated features // BMC Med. Genomics. – 2018. – Vol. 11(Suppl 6). – Art. ID 120. https://doi.org/10.1186/s12920-018-0436-9;

11. Gu Q., Feng Z., Liang Q., Li M., Deng J.., Ma M, Wang W.., Liu J., Liu P, Rong P. Machine learning-based radiomics strategy for prediction of cell proliferation in non-small cell lung cancer // Eur. J. Radiol. – 2019. – Vol. 118. – P. 32-37. https://doi.org/10.1016/j.ejrad.2019.06.025;

12. Bergquist S.L., Brooks G.A., Keating N.L., Landrum M.B., Rose S. Classifying Lung Cancer Severity with Ensemble Machine Learning in Health Care Claims Data // Proc. Mach. Learn. Res. – 2017. – Vol. 68. – P. 25-38. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6287925/;

13. Sena G.R., Lima T.P.F., Mello M.J.G., Thuler L.C.S., Lima J.T.O. Developing Machine Learning Algorithms for the Prediction of Early Death in Elderly Cancer Patients: Usability Study // JMIR Cancer. – 2019. – Vol. 5(2). – Art. ID e12163. https://doi.org/10.2196/12163;

14. Siah K.W., Khozin S., Wong C.H., Lo A.W. Machine-Learning and Stochastic Tumor Growth Models for Predicting Outcomes in Patients With Advanced Non-Small-Cell Lung Cancer // JCO Clin Cancer Inform. – 2019. – Vol. 3. – P. 1-11. https://doi.org/10.1200/CCI.19.00046.

ТҰЖЫРЫМ

ҚАЗАҚСТАН РЕСПУБЛИКАСЫНДАҒЫ ӨКПЕНІҢ ҚАТЕРЛІ ІСІГІМЕН АУЫРАТЫН НАУҚАСТАРДЫҢ ӨМІР СҮРУ КӨРСЕТКІШІ НӘТИЖЕЛЕРІН БОЛЖАМДАУ МОДЕЛІН ҚҰРАСТЫРУДАҒЫ МАШИНАМЕН ОҚЫТУДЫҢ РӨЛІ

В.А. Макаров[1,2], Д.Р. Кайдарова[3], С.Е. Есентаева[4], Ж. Калматаева[2], М.Е. Мансурова[2], Н. Кадырбек[2], Р.Е. Кадырбаева[3], С.Т. Олжаев[1], И.И. Новиков[1]

[1]«Алматы Жергілікті Көпсалалы Клиникасы» КМК ШЖҚ, Алматы, Қазақстан Республикасы;
[2]«Әл-Фараби атындағы Қазақ Ұлттық университеті» КеАҚ, Алматы, Қазақстан Республикасы;
[3]«Қазақ онкология және радиология ғылыми-зерттеу институты» АҚ, Алматы, Қазақстан Республикасы;
[4]"Қазақстан-Ресей медициналық университеті» ҰҚУ, Алматы, Қазақстан Республикасы

Өзектілігі: *Іа сатысындағы өкпенің қатерлі ісігінің 5 жылдық жалпы өмір сүру деңгейі 73% құрайды, ал радикалды емделген пациенттерде рецидив жиілігі шамамен 10% құрайды.*

Зерттеу мақсаты *– бірқатар клиникалық және морфологиялық факторлардың болжамды маңыздылығы мен өкпе қатерлі ісігі бар науқастардың жалпы өмір сүру нәтижелерін болжау үшін машиналық оқыту алгоритмдерін қолдану мүмкіншілігін бағалау.*

Әдістер: *030-6/у с34 – өкпе обыры (n=19379) нысандарына 2014-2018 жж. ОНЭР деректер базасынан талдау жүргізілді, Каплан-Мейер әдісі бойынша жалпы өмір сүруге қауіп факторларының әсерін бағалау жүргізілді. Тиісінше, болжау модельдерін құруға арналған оқыту жиынтығы 19379 бақылау мен 15 факторды қамтиды. Жұмыста қолданылатын Машиналық оқыту алгоритмдері (Random Forest Classifier, Gradient Boosting Classifier, Logistic Regression Model, Decision Tree Classifier, K Nearest Neighbors (KNN) Classifier) Python*

бағдарламалау тілінде іске асырылған. Нәтижелер қате матрицаны құру, жіктеу өлшемдерін есептеу арқылы бағаланды: оқыту және тексеру (validation), дәлдік (precision), толықтық (recall), Каппа-Коэн кезінде дұрыс жіктелген объектілердің үлесі (accuracy).

Нәтижелері: *Біздің зерттеуімізде 19 379 науқас талданды, оның ішінде 15 494 ер адам (79,95%) және 3 885 әйел (20,04%). Зерттеу барысында қазіргі күні ерлер арасында 6 171 науқас (39,8%) тірі екендігі анықталды, бұл ретте өмір сүру медианасы 8,3 айды құрады (SE – 0,154 ай, 95% ДИ – 7,96-8,56). Әйелдер арасында 1 962 науқас (49,5%) тірі, бұл ретте өмір сүру медианасы 15,43 айды құрады (SE – 1,0 ай, 95% ДИ – 13,497-17,363). 1 037 пациентте (5,35%) аурудың I сатысында және 4 145 (21,38%) II сатысында анықталды. ӨҰЖЕҚІ науқастардың көпшілігінде (61,4%) кең таралған сатыда диагноз қойылған: 9 189 адамда (47,4%) – III сатыда, 4 655-те (24%) – IV сатыда. Өмір сүру медианасындағы айырмашылықтардың дұрыстығын бағалау ($\chi2$=3991,6, p=0,00) ісік процесінің болжамды маңыздылығын және науқастардың өмір сүруіне әсерін көрсетеді. Сондай-ақ, өкпе қатерлі ісігінің әртүрлі морфологиялық формалары бар науқастар арасында өмір сүру медианасындағы айтарлықтай айырмашылық морфологиялық фактордың болжамды маңыздылығы туралы айтуға мүмкіндік береді (статистикалық тұрғыдан алғанда, бұл көрсеткіштер арасындағы айырмашылық сенімді болды, $\chi2$=623,4 p=0,000).*

Қорытынды: *Машиналық оқыту модельдері хирургиялық емдеуден кейін де, ОНЭР дерекқорына тіркелгеннен кейін де науқастардың өлім қаупін болжауға мүмкіндік береді. Науқасқа бағдарланған медициналық шешімдер қабылдауды қолдау жүйесін құру адъювантты терапияның, диспансерлік бақылаудың және диагностикалық зерттеулер жиілігінің оңтайлы стратегияларын таңдауға мүмкіндік береді.*

Түйінді сөздер: *өкпенің қатерлі ісігі, болжамды маңыздылығы, машиналық оқыту, қайталанулар, жалпы өмір сүру көрсеткіші.*

ABSTRACT

THE ROLE OF MACHINE LEARNING IN THE DEVELOPMENT OF A MODEL FOR PREDICTING THE SURVIVAL OF LUNG CANCER PATIENTS IN THE REPUBLIC OF KAZAKHSTAN

V.A. Makarov[1,2], D.R. Kaidarova[3], S.E. Yessentayeva[4], J. Kalmatayeva[2], M.E. Mansurova[2], N. Kadyrbek[2], R.E. Kadyrbayeva[3], S.T. Olzhayev[1], I.I. Novikov[1]

[1]«Almaty Regional Multidisciplinary Clinic» MSE on REM, Almaty, the Republic of Kazakhstan;;
[2]«Al-Farabi Kazakh National University» Non-Commercial JSC, Almaty, the Republic of Kazakhstan;
[3]«Kazakh Institute of Oncology and Radiology» JSC, Almaty, the Republic of Kazakhstan;
[4]«Kazakh-Russian Medical University» Non-Governmental Educational Institution, Almaty, the Republic of Kazakhstan

Relevance: *The 5-year overall survival rate(s) in NSCLC p-stage IA is 73%, and the recurrence rate in radically treated patients is almost 10%.*

The study aimed to *evaluate the prognostic significance of several clinical and morphological factors and apply machine learning algorithms to predict the results of overall survival of patients with lung cancer.*

Methods: *The forms 030-6/y C34 – lung cancer (n=19,379) from the EROB database for 2014-2018 were analyzed, and the impact of risk factors on overall survival was assessed using the Kaplan-Meier method. Accordingly, the training data set for constructing forecasting models included 19,379 observations and 15 factors. The machine learning algorithms such as Random Forest Classifier, Gradient Boosting Classifier, Logistic Regression Model, Decision Tree Classifier, and K Nearest Neighbors (KNN) Classifier were implemented in the Python programming language. The results were evaluated by constructing an error matrix, calculating classification metrics: the proportion of correctly classified objects (accuracy) during training and validation (validation), accuracy (precision), completeness (recall), Kappa-Cohen.*

Results: *In our study, 19,379 patients were analyzed, including 15,494 men (79.95%) and 3,885 women (20.04%). At the time of the study, 6,171 men (39.8%) and 1,962 women (49.5%) were alive. Median survival was 8.3 months (SE – 0.154 months, 95% CI – 7.96-8.56) in men and 15.43 months (SE – 1.0 months, 95% CI – 13.497-17.363) in women. At diagnosis, 1,037 patients (5.35%) had stage I disease, other 4,145 (21.38%) had stage II. Most patients (61.4%) had advanced stage NSCLC: 9,189 people (47.4%) were diagnosed with stage III, and 4,655 (24%) – with stage IV. The reliability of differences in median survival ($\chi2$=3991.6, p=0.00) indicated the prognostic significance of the tumor process stage and its influence on the patient's survival. Also, the revealed significant difference in the median survival of patients with various morphological forms of lung cancer suggests the prognostic significance of the morphological factor (the difference between those indicators was statistically significant, $\chi2$=623.4 p=0.000).*

Conclusion: *Machine learning models can predict the risk of fatal outcomes for patients after surgical treatment and registration in the EROB database. The creation of patient-oriented systems to support medical decision-making makes it possible to choose the optimal strategies for adjuvant therapy, dispensary observation, and frequency of diagnostic studies.*

Keywords: *lung cancer, prognostic significance, machine learning, relapses, overall survival.*

Прозрачность исследования: *Авторы несут полную ответственность за содержание данной статьи.*
Конфликт интересов: *Авторы заявляют об отсутствии конфликта интересов.*
Вклад авторов: *вклад в концепцию – Макаров В.А., Кайдарова Д.Р., Есентаева С.Е., Олжаев С.Т., Новиков И.И.; научный дизайн – Макаров В.А., Есентаева С.Е., Калматаева Ж.А.; исполнение заявленного научного исследования – Макаров В.А., Мансурова М.Е., Кадырбек Н., Кадырбаева Р.Е.; интерпретация заявленного научного исследования – Макаров В.А., Мансурова М.Е., Кадырбек Н.; создание научной статьи – Макаров В.А., Мансурова М.Е., Кадырбек Н., Кадырбаева Р.Е.*

Сведения об авторах:
Макаров Валерий Анатольевич – зав. хирургическим отделением, Алматинская Региональная Многопрофильная Клиника, Алматы, Республика Казахстан, тел: +77017750830; e-mail: makaroff_valeriy@mail.ru, ID ORCID: https://orcid.org/0000-0003-2120-5323;
Кайдарова Диляра Радиковна – д.м.н., профессор, академик НАН РК, председатель Правления АО «КазНИИОиР», президент Ассоциации онкологов и радиологов РК, Алматы, Республика Казахстан, тел: +77017116593; e-mail: kazior@onco.kz, ID ORCID: https://orcid.org/0000-0003-2120-5104;
Есентаева Сурия Ертугыровна – д.м.н., профессор, зав. кафедрой онкологии, Казахско-Российский Медицинский Университет, Алматы, Республика Казахстан, тел: +77077942910, e-mail: surya_esentay@mail.ru, ID ORCID: https://orcid.org/0000-0001-7087-1440;
Калматаева Жанна – д.м.н., профессор, декан факультета медицины и общественного здравоохранения, Казахский Национальный Университет им. аль-Фараби, Алматы, Республика Казахстан, тел: +77772187666, e-mail: Zhanna.Kalmatayeva@kaznu.kz, ID ORCID: https://orcid.org/0000-0002-5562-1969;
Мансурова Мадина Есимхановна – преподаватель, Казахский Национальный Университет им. аль-Фараби, Алматы, Республика Казахстан, тел: + 77014151960, e-mail: madina.mansurova@kaznu.kz, ID ORCID: https://orcid.org/0000-0002-9680-2758;
Кадырбек Нургали – преподаватель, Казахский Национальный Университет им. аль-Фараби, Алматы, Республика Казахстан, тел: +77079890989, e-mail: mailto:mnurgaliqadyrbek@gmail.com, ID ORCID: https://orcid.org/0000-0002-5461-8899;
Кадырбаева Рабига Есенгаликызы (корреспондирующий автор) – химиотерапевт, АО «Казахский Научно-Исследовательский Институт Онкологии и Радиологии», Алматы, Республика Казахстан, тел: +77074023344; e-mail: rabiga-92@mail.ru, ID ORCID: https://orcid.org/0000-0001-8254-8675;
Олжаев Саяхат Турарбекович – директор Алматинской региональной многопрофильной клиники, Алматы, Республика Казахстан, тел: +77017749999, e-mail: S.Olzhayev20@gmail.com, ID ORCID: https://orcid.org/0000-0002-3312-323X;
Новиков Игорь Игоревич – заместитель директора по лечебной части, Алматинская Региональная Многопрофильная Клиника, Алматы, Республика Казахстан, тел: +77773640684, e-mail: mailto:migor-novikov-1982@mail.ru, ID ORCID: https://orcid.org/0000-0001-7015-6770.