

ПРОГНОЗИРОВАНИЕ ТЕЧЕНИЯ ОНКОЛОГИЧЕСКОГО ЗАБОЛЕВАНИЯ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Шепетько М.Н., Карапетян Г.М., Ермаков В.В., Косик И.И.

Учреждение образования «Белорусский государственный медицинский университет»,
Минск, Республика Беларусь
e-mail: greg.itlab@gmail.com

Аннотация: *совершенствование клинических протоколов, применяемых при лечении онкологических пациентов, в значительной степени зависит от повышения точности прогноза за счёт его большей персонификации. Предлагается вариант метода персонифицированного прогнозирования течения рака лёгкого, основанный на возможностях модели с машинным обучением. Для создания модели использовался алгоритм «классификатор деревьев» с применением градиентного спуска библиотеки CatBoost. Чувствительность метода 78,8%, специфичность 88,8%.*

Ключевые слова: *рак лёгкого, искусственный интеллект, машинное обучение, персонифицированный прогноз, функция дожития.*

USING MACHINE LEARNING METHODS TO PREDICT THE COURSE OF CANCER

Shapetska M.N., Karapetsian R.M., Ermakov V.V., Kosik I.I.

Belarussian State Medical University Minsk, Republic of Belarus

e-mail: greg.itlab@gmail.com

Abstract. Improving clinical protocols used for treating cancer patients largely depends on increasing the accuracy of prognosis by making it more personalized. A variant of a personalized method for predicting the course of lung cancer is proposed, based on the capabilities of a machine learning model. The “tree classifier” algorithm with gradient descent from the CatBoost library was used to create the model. The sensitivity of the method is 78,8%, the specificity is 88,8%.

Keywords: lung cancer, artificial intelligence, machine learning, personalized prognosis, survival function.

Введение. Применяемые в клинической практике протоколы лечения онкобольных разрабатываются с учётом широкого набора клинических и анамнестических показателей. Важные позиции при этом отводятся показателям, на основе которых становится возможным формирование прогноза течения заболевания. Таким как стадия рака, определяемая согласно TNM-классификации, морфологический тип злокачественного новообразования, возраст пациента, его пол, наличие сопутствующих заболеваний, генетические предикторы и т.д. Спектр показателей настолько широк, что выявить все значимые сочетания, а затем определить их влияние на развитие и исход заболевания в каждом конкретном случае очень сложно. Отсюда присутствие и в прогнозах, и в клинических рекомендациях определённой степени усреднённости. В связи с этим совершенствование клинических протоколов в значительной степени зависит от повышения точности прогноза за счёт его большей персонификации. Одним из действенных приёмов в создании прогностических моделей, способных выявлять скрытые закономерности при анализе большого объёма данных, является использование возможностей искусственного интеллекта.

Цель работы. С применением методов машинного обучения предложить модель прогнозирования вероятного течения заболевания у пациентов с установленным диагнозом рак лёгкого. В качестве объекта исследования выбран именно рак лёгкого по двум причинам. Во-первых, потому что среди пациентов этого профиля показатель 5-летней выживаемости на протяжении последних десятилетий не превышает 20-25% [1] даже для радикально оперированных, что обуславливает необходимость получения максимально точного прогноза для своевременной корректировки объёмов противоопухолевой терапии. Во-вторых, поскольку в перечень анализируемых данных кроме анамнестических и клинических решено было включить генетические показатели пациентов, а молекулярно-генетическая природа этого вида онкологического заболевания до настоящего времени недостаточно исследована.

Материалы и методы. В исследовании использовалась база с ретропроспективными данными по 635 пациентам, проходившим лечение в городском клиническом онкологическом диспансере Минска. База данных состояла из совокупного набора 45-ти анамнестических (пол, возраст,

профессия, онкоанамнез, статус курения и т.д.), клинических (наличие и вид операции, гистологический тип опухоли, локализация опухоли, размеры опухоли, стадия заболевания, сопутствующие заболевания и т.д.), молекулярно-генетических показателей (полиморфные варианты генов, ответственных за пролиферацию, дифференциацию, апоптоз, детоксикацию и т.д.). Кроме этого в обязательном порядке по каждому пациенту присутствовал параметр «исход заболевания», который мог принимать одно из двух значений – пациент умер или пациент жив. Набор показателей формировался на основании собственного опыта хирургического лечения и с учётом данных аналогичных зарубежных исследований, посвящённых лечению рака лёгких. Под раком лёгкого подразумевается собирательное понятие, объединяющее злокачественные образования в лёгких с разной гистологической структурой. Генетический анализ проводился в Институте генетики и цитологии НАН Беларуси.

База данных создавалась с целью последующего обучения прогностической модели. Однако во избежание т.н. эффекта переобучения (когда из-за избытка не значимой информации модель начинает больше реагировать на лишние, незначительные детали, «снижая интерес» к реальным закономерностям), из полного набора показателей следовало сначала выбрать те, от которых достоверно зависело значение параметра «исход заболевания». Предварительная обработка данных осуществлялась с применением методов математической статистики. Для анализа зависимости исхода заболевания от того или иного показателя использовался Хи-квадрат критерий Пирсона. Сравнивались, например, частоты встречаемости неблагоприятного исхода среди курящих и некурящих пациентов. При небольших объемах выборки применялся Критерий Фишера (ϕ). Значимость того или иного показателя определялась в ходе оценки степени его влияния на известный исход заболевания. Для этого с помощью коэффициента сопряжённости Чупрова (C) и критерия VКрамера определялась сила связи исходов с показателями. В итоге из начального набора, предназначенного для обучения прогностических моделей, были исключены показатели, влияние которых на исход заболевания не удалось подтвердить, либо оно оказалось опосредованным. Так, показатель «Локализация опухоли» был исключён из обучающей выборки, поскольку обнаружена существенная корреляция (в пределах 35%) между локализацией опухоли и её гистологическим типом. Анализ показал, что у пациентов, объединённых в группы по гистологическому диагнозу, опухоли центральной локализации чаще принадлежат плоскоклеточному и мелкоклеточному раку. А поскольку именно по этим вариантам обнаружился наибольший уровень летальности, было принято предположение о том, что высокий процент неблагоприятных исходов при опухолях центральной локализации обусловлен прежде всего корреляцией с гистологическим диагнозом. Т.е., влияние фактора локализации опухоли на исход заболевания носит опосредованный характер и во избежание перегрузки модели прогнозирования большим количеством входных параметров, этим показателем имеет смысл пренебречь.

В процессе обучения модели список обучающих параметров дополнительно корректировался в сторону его сокращения. Для этого

использовались весовые коэффициенты каждого из параметров, рассчитываемые самой прогностической моделью. Обучение и тестирование проводилось на данных пациентов, у которых количество пропусков значений по отобранным показателям сводилось к разумному минимуму. Из 390 таких пациентов в тренировочную выборку вошли 260, в валидационную 130 человек. В конечном итоге были выделены сначала 20, а затем 11 показателей с наибольшими значениями весовых коэффициентов. В список наиболее информативных вошли те, средние значения (за 50 итераций обучения модели) весовых коэффициентов которых расположились на одиннадцати первых позициях. Обучение проводилось на данных пациентов, случайным образом отобранных из тренировочной группы для каждой итерации. Итоговый список включал в себя следующие показатели (каждый из которых мог принимать от 2 до 5 значений): возраст, семейный онкоанамнез, статус курения, наличие сопутствующих заболеваний, стадия заболевания, операция, гистологический диагноз, степень дифференцировки опухоли, размер опухоли, гены. В показатель «гены» вошли: ген VEGF, ген NAT, ген GSTP, ген MMP, ген DNMT, ген-рецептор KDR. Для гена VEGF прогностическая модель особо выделила комбинацию *VEGF 634 G/C - VEGF 936 C/C – VEG 2578* (с любым вариантом полиморфизма). Как видно, неблагоприятный прогноз в значительной степени формируется сбоями в работе генов, отвечающих за систему детоксикации, образование сосудистой сети и пролиферацию клеток. Два последних вида взаимоувязаны, поскольку интенсивное кровоснабжение стимулирует клеточную активность.

Предварительно для создания прогностической модели обучались и тестировались несколько классических алгоритмов с машинным обучением: классификатор случайных деревьев, логистическая регрессия, ElasticNet [2], регрессия Кокса, классификатор деревьев с применением градиентного спуска библиотеки CatBoost [3]. Последняя из перечисленных моделей показала наибольший по точности результат и была выбрана в качестве основной модели. Выбор модели, однако, был обусловлен не только значением точности, но и рядом других преимуществ. Модель деревьев решений, созданная с применением библиотеки CatBoost, использует градиентный бустинг для обучения. При таком обучении ансамбль деревьев решений создаётся последовательно, и каждое последующее дерево, обучаясь на корректировке ансамбля, уменьшает его ошибку. То есть, каждое последующее дерево улучшает общий результат ансамбля деревьев. Подобный подход позволяет находить сложные нелинейные зависимости между данными. Ещё одним преимуществом использования CatBoost является встроенная работа с гетерогенными данными, в том числе категориальными (например, пол пациента или гистологический диагноз) - данными, не обязанными иметь численное значение. Способность библиотеки анализировать параметры модели, делает некритичным пропуски их значений.

Кроме всего перечисленного важным аргументом в пользу CatBoost стала предоставляемая в ней возможность выбора функции ошибок для обучения модели. В качестве такой функции была выбрана функция ошибок модели пропорциональных рисков Кокса (Cox) [4].

Функции, позволяющей прогнозировать по времени риск наступления события, зависящего от набора независимых предикторов. Построенная и обученная таким образом модель определяет вероятность наступления события (в нашем случае неблагоприятного исхода), как функцию времени, поэтому для определения точности её предсказания в качестве метрики использовался индекс соответствия C-index (concordance index) [5]. Эта метрика оценивает не только полученное моделью значение вероятности исхода, но и определённый ею временной интервал для этого значения. Т.е., для метрики важнее не то, что модель правильно указала выживших и умерших, а то, что она правильно указала какой пациент умер раньше, какой позже. Метрика может принимать значения от 1 - абсолютно правильное предсказания по всем пациентам, до 0 - абсолютно неправильное предсказание. Точность предсказания выбранной прогностической модели оказалась равной 0,76.

Формируемое моделью значение опасности наступления события $M(x, \beta)$ используется для получения кумулятивной функции риска (1):

$$H(t|x) = H_0(t) \cdot e^{M(x, \beta)} \quad (1),$$
 где H_0 – базовая кумулятивная функция опасности, рассчитываемая по методу Бреслау (Breslow) и характеризующая некое усреднённое значение опасности неблагоприятного исхода в каждый момент времени, рассчитанное для группы пациентов.

$M(x, \beta)$ – значение опасности, которое формирует используемая нами модель.

x – предикторы, оказывающие влияние на риск наступления события (в нашем случае – неблагоприятного исхода заболевания).

β – регрессионные коэффициенты предикторов.

Кумулятивные функции риска $H(t|x)$ и $H_0(t)$ необходимы для получения с помощью формулы (2) т.н. функции дожития $S(t)$ – индивидуальной или базовой (для группы пациентов):

$$S(t) = e^{-H(t)} \quad (2)$$

Результаты и обсуждение. Рассчитанная по формуле (2) функция дожития $S(t)$ описывает вероятность благоприятного течения заболевания на произвольно выбранном отрезке времени. Чем круче кривая уходит вниз, тем хуже прогноз (рис.1). На рисунке 1 демонстрируются индивидуальные кривые дожития, построенные для четырёх пациентов с использованием их кумулятивных функций риска $H(t|x)$. Худший прогноз демонстрирует оранжевая кривая, принадлежащая пациенту №147.

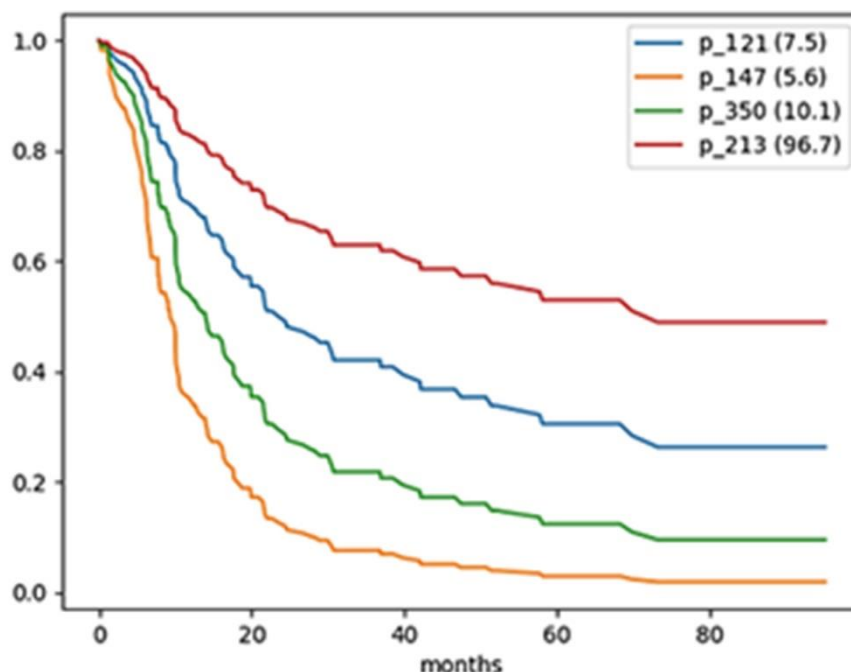


Рисунок 1 – Кривые дожития 4-х пациентов. Ось ординат – значения вероятности дожития в любой момент времени, выбранный на оси абсцисс

График функции $S(t)$ позволяет получить наглядную интерпретацию прогноза, если сопоставлять индивидуальные кривые дожития конкретных пациентов не между собой, а с базовой кривой, характеризующей некое усреднённое течение заболевания для пациентов, объединённых в группу по какому-то существенному признаку рассматриваемого заболевания, например, по его стадии. Такой подход позволяет увидеть, является ли риск неблагоприятного исхода для конкретного пациента выше или ниже риска по группе. Но при этом требуется заменить общую прогностическую модель на четыре модели, каждую из которых необходимо переобучить на данных пациентов с соответствующей стадией заболевания. Дополнительно для исследования важности показателя «операция» из пациентов с четвертой стадией была создана и обучена выборка только прооперированных пациентов. Точность полученных моделей изменилась в основном в меньшую сторону, поскольку после перегруппировки пациентов уменьшились объёмы тренировочных выборок. Точность предсказания для пациентов с первой стадией уменьшилась до 0,65. Второй и третьей до 0,7. Четвёртой, наоборот, возросла до 0,86. Однако, несмотря на это, ценность прогноза и его адекватность повысились. Основным критерием для формулирования прогностического заключения стало положение кривой дожития пациента с базовыми кривыми, полученными для четырёх стадий рака лёгкого. Метод прогнозирования, основанный на применении разработанной модели, предполагает формирование прогностического заключения и рекомендаций по следующему алгоритму:

а) если кривая дожития совпадает или находится выше базовой кривой для стадии, соответствующей стадии заболевания пациента, или не пересекается с базовой кривой стадии, превышающей стадию заболевания пациента – прогноз благоприятный. Лечение проводится в полном соответствии с принятым для данной стадии клиническим протоколом лечения.

б) если кривая выживаемости хотя бы в одной точке пересекается с кривой, которая является базовой для стадии, превышающей стадию заболевания пациента – прогноз неблагоприятный. К нему рекомендуется применить клинический протокол лечения, соответствующий более высокой стадии опухолевого процесса, или индивидуализировать противоопухолевое лечение в соответствии с алгоритмами диагностики и лечения пациентов со злокачественными новообразованиями вследствие того, что пациент попадает в зону неблагоприятного прогноза.

На рисунке 2 приведён пример анализа вероятного течения рака лёгкого у пациента «А» 1949 года рождения с первой стадией заболевания.

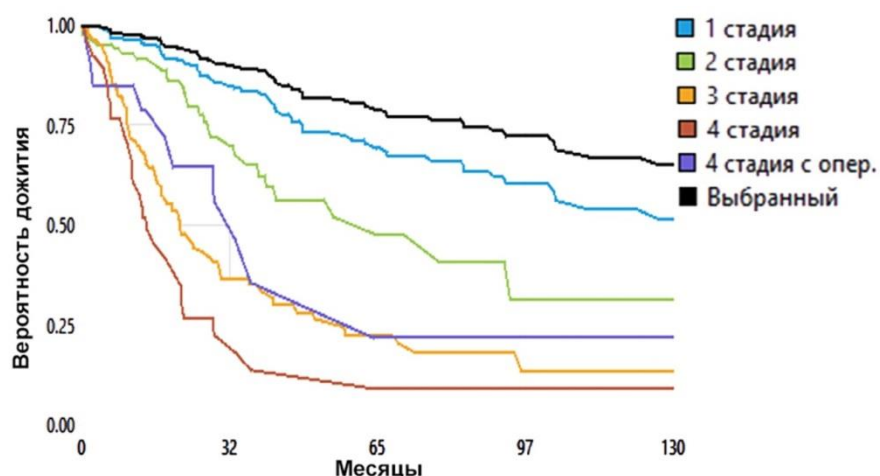


Рисунок 2 – График дожития. Кривая дожития пациента «А» чёрного цвета. Построенные для каждой стадии базовые кривые имеют собственные цвета.

Как видно на рисунке 2, кривая дожития пациента находится выше голубой базовой кривой, описывающей характерное течение заболевания у пациентов с аналогичной стадией, что позволяет прогнозировать отсутствие в дальнейшем непредвиденных осложнений. Таким образом, пациент «А» не нуждается в коррекции стандартного протокола лечения, применяемого для пациентов с первой стадией рака лёгкого.

Следует так же обратить внимание на базовую кривую (синего цвета), построенную моделью для прооперированных пациентов 4-ой стадии, выделенных в дополнительную группу. Её расположение указывает на важность своевременного хирургического вмешательства. Прогноз течения заболевания прооперированных пациентов с 4-ой стадией значительно благоприятнее прогноза для смешанной группы этой же стадии (коричневая кривая) и в целом даже лучше прогноза для пациентов менее опасной, но также смешанной группы с 3-ей стадией рака лёгкого (оранжевая кривая).

Пример выраженного неблагоприятного прогноза, построенного для пациента «Б», 1952 года рождения, со второй стадией рака лёгкого, демонстрируется на рисунке 3.

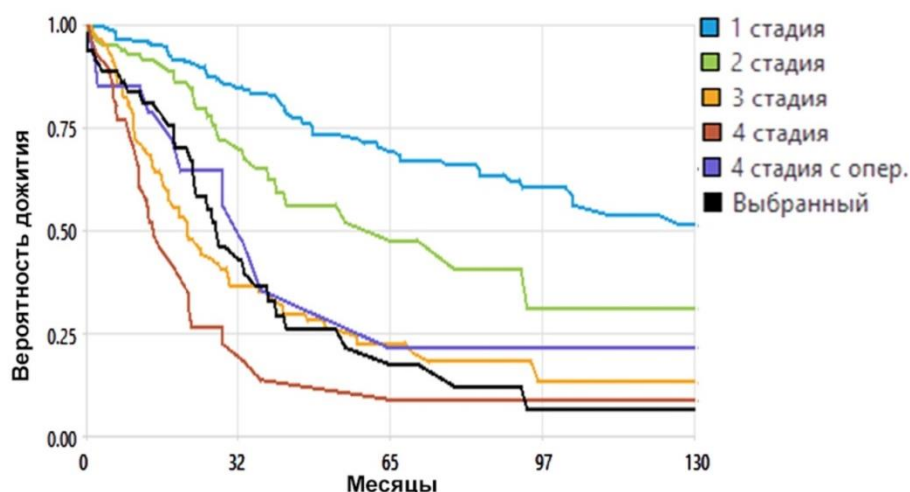


Рисунок 3 – График дожития. Кривая дожития пациента «Б» чёрного цвета.

Построенные для каждой стадии базовые кривые имеют собственные цвета

Кривая дожития пациента «Б», построенная прогностической моделью, с первых месяцев резко уходит вниз, пересекаясь с базовыми кривыми третьей, а в последующем и четвёртой стадии. Это свидетельствует о резко негативном прогнозе и необходимости применения противоопухолевой терапии сверх той, что предусмотрена клиническим протоколом для второй стадии данного заболевания.

Для тестирования метода прогнозирования исхода рака лёгкого использовались данные 160 пациентов, разделённых на две равные группы по показателю «исход заболевания». Периоды наблюдения пациентов с благоприятным исходом были разные, но не менее трёх лет. В группу с неблагоприятным исходом отбирались пациенты с летальным исходом, зафиксированным в первые пять лет наблюдения. Обе группы в свою очередь были разделены на 4 подгруппы, в каждую из которых отбирались по 20 пациентов с одинаковой стадией рака лёгкого. Задача основанного на прогностической модели метода прогнозирования – выявить среди пациентов с установленным диагнозом тех, которые нуждаются в усилении противоопухолевой терапии. Поэтому группу, составленную из тестируемых с неблагоприятным исходом, условно назовём «экспериментальной», а группу пациентов, живых на момент тестирования, обозначим как «контрольную». Тогда полученные для контрольной группы результаты (таблица 1) характеризуют специфичность предложенного метода прогнозирования.

О чувствительности метода прогнозирования можно судить по результатам, полученным для экспериментальной группы (таблица 2).

Таблица 1

Результаты тестирования пациентов контрольной группы

Стадии	Количество пациентов	Верное заключение модели	Неверное заключение модели	Специфичность для стадии %
I	20	18	2	90
II	20	20	-	100

III	20	19	1	95
IV	20	14	6	70

Представленные в таблице 1 данные свидетельствуют о том, что в целом из 80 пациентов с разными стадиями для девяти пациентов был дан неверный прогноз на неблагоприятный исход. Следовательно, для всей контрольной выборки специфичность алгоритма оказалась равной 88,8%. Но если смотреть специфичность по стадиям, то видно, что по трём первым стадиям специфичность в целом составила 95%. Т.е., предложенный алгоритм с достаточно высокой точностью не предложит повышения интенсивности терапии без реальной необходимости. Наибольшее количество неточных прогнозов пришлось на группу пациентов с самой тяжёлой – четвёртой стадией, к которым априори применялись самые активные методы лечения. Таким, образом алгоритм в любом случае не ухудшит стандартный протокол лечения.

Таблица 2

Результаты тестирования пациентов экспериментальной группы

Стадии	Количество пациентов	Верное заключение модели	Неверное заключение модели	Чувствительность для стадии %
I	20	13	7	65
II	20	15	5	75
III	20	15	5	75
IV	20	20	-	100

Из представленных в таблице 2 данных можно сделать вывод, что в целом из 80 пациентов с разными стадиями алгоритм дал верный прогноз на неблагоприятный исход для 63 пациентов. Т.е. для всей контрольной выборки чувствительность алгоритма оказалась равной 78,8%. Если убрать из анализа пациентов с 4-ой стадией, для которых сделать лечение более интенсивным, чем уже предпринимаемое, невозможно, то чувствительность метода окажется равной 71,8%. Т.е., в случае применения алгоритма, у каждого 7 из 10 умерших пациентов появился бы шанс на увеличение продолжительности жизни за счёт более интенсивного лечения, на необходимость которого указывает полученный алгоритмом неблагоприятный прогноз.

Предложенный метод предоставляет возможность получать персонализированный прогноз течения рака лёгкого и таким образом выявлять пациентов, нуждающихся в коррекции противоопухолевого лечения вследствие высокого риска неблагоприятного исхода заболевания. Индивидуальная коррекция протокола лечения способствует росту результативности терапии и увеличению продолжительности жизни пациента.

Список литературы

1. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries / F. Bray [et al.] // A Cancer Journal for Clinicians. – 2018. – Vol. 68, № 6. – P. 394-424.
2. Zou, H. Regularization and variable selection via the elastic net / H. Zou, T. Hastie // Journal of the Royal Statistical Society Series B: Statistical Methodology. – 2005. – Vol. 67, № 2. – P. 301-320.
3. Cat Boost: unbiased boosting with categorical features / L. Prokhorenkova, G. Gusev, A. Vorobev [et al.] // arXiv preprint arXiv: 1706.09516v5. – 2019. – 23 p.
4. Cox, D.R. Regression Models and Life-Tables // Journal of the Royal Statistical Society Series B: Statistical Methodology. – 1972. – Vol. 34, № 2. – P. 187-220.
5. Prioritized Concordance Index for Hierarchical Survival Outcomes / C. Li, Qing Pan, Noorie Hyun [et al.] // Stat. Med. – 2019. – Vol. 10. – № 38(15). – P. 2868–2882.

Федеральное государственное бюджетное
образовательное учреждение
Кубанский государственный медицинский университет
Министерства здравоохранения Российской Федерации
ФГБОУ ВО КубГМУ Минздрава России



Материалы

Четырнадцатой международной учебно-методической конференции
«Инновации в образовании»



г. Краснодар, 11 апреля 2024 года