Оценка риска неблагоприятного клинического исхода методами углубленного анализа данных

Б. Э. Горный[1], А. П. Рыжов[2], А. С. Строгалов[3], А. Д. Журавлев[4], А. А. Хусаенов[5], И. А. Шергин[6], Д. А. Фещенко[7], А. М. Абдуллаев[8], А. В. Концевая[9]

[1] *Горный Борис Эмануилович* — к.м.н, ведущий научный сотрудник отдела первичной профилактики ХНИЗ в системе здравоохранения ФГБУ "НМИЦТПМ", e-mail: bgornyy@gnicpm.ru

Gornyi Boris Emanuilovich — Ph.D. in medical science, leading reseacher of department of Primary Prevention of Chronic Non-Communicable Diseases in the Healthcare System, National Medical Research Center for Therapy and Preventive Medicine

[2] *Рыжов Александр Павлович* — д.т.н., к.ф.-м.н., MBA, профессор кафедры математической теории интеллектуальных систем мех.-мат. ф-та МГУ; e-mail: alexander.ryjov@gmail.com

Ryjov Alexander Pavlovich — Sc.D. in engineering, Ph.D. in mathematics, MBA, professor, Moscow State University, Faculty of Mechanics and Mathematics, Mathematical Theory of Intelligent Systems department

[3] *Строгалов Александр Сергеевич* — к.ф.-м.н., доцент кафедры математической теории интеллектуальных систем мех.-мат. ф-та МГУ; e-mail: strogalov@mail.ru

Strogalov Alexander Sergeevich — Ph.D. in mathematics, associate professor, Moscow State University, faculty of Mechanics and Mathematics, Mathematical Theory of Intelligent Systems department

[4] *Журавлев Артем Дмитриевич* — аспирант кафедры математической теории интеллектуальных систем мех.-мат. ф-та МГУ; e-mail: artemzhuravlev.msu@gmail.com

Zhuravlev Artem Dmitrievich — postgraduate student, Moscow State University, faculty of Mechanics and Mathematics, Mathematical Theory of Intelligent Systems department

[5] *Хусаенов Артем Азатович* — аспирант кафедры математической теории интеллектуальных систем мех.-мат. ф-та МГУ; e-mail: a.khusaenov@mail.ru

Khusaenov Artem Azatovich — postgraduate student, Moscow State University, faculty of Mechanics and Mathematics, Mathematical Theory of Intelligent Systems department

[6] *Шергин Иван Андреевич* — студент 5 курса кафедры математической теории интеллектуальных систем мех.-мат. ф-та МГУ; e-mail: i.a.shergin@gmail.com

Shergin Ivan Andreevich — 5 year student, Moscow State University, faculty of Mechanics and Mathematics, Mathematical Theory of Intelligent Systems department

[7] *Фещенко Дарья Анатольевна* — заведующая операционным блоком ФГБУ "НМИЦТПМ"; e-mail: dfeshenko@gnicpm.ru

Feshchenko Daria Anatol'evna — the head of surgical block of National Medical Research Center for Therapy and Preventive Medicine, Moscow

[8] *Абдуллаев Алсан Мурадович* — аспирант отдела нарушений сердечного ритма и проводимости ФГБУ "НМИЦТПМ"; e-mail: abdullaevaslanm@mail.ru

Abdullaev Aslan Muradovich, postgraduate student, National Medical Research Center for Therapy and Preventive Medicine, Moscow

[9] *Концевая Анна Васильевна* — д.м.н, заместитель директора ФГБУ "НМИЦТПМ"; e-mail: AKontsevaya@gnicpm.ru

Kontsevaya Anna Vasilievna, Doctor of Medicine, deputy director of National Medical Research Center for Therapy and Preventive Medicine, Moscow

Возникновение неблагоприятных событий в процессе оказания медицинской помощи возникает у 10-15% госпитализированных пациентов. Снижение даже на несколько процентов возникновения таких событий позволит сохранить тысячи жизней. Одним из путей решения этой важнейшей проблемы является использование интеллектуальных информационных технологий, позволяющих прогнозировать риск возникновения неблагоприятного клинического исхода у пациентов. В работе представлены результаты исследования [1], выполненного совместно сотрудниками Национального медицинского исследовательского центра терапии и профилактической медицины МЗ РФ и механико-математического факультета МГУ имени М.В. Ломоносова, показывающего применимость методов анализа данных в решении этой важной проблемы.

Ключевые слова: профилактическая медицина, неблагоприятный клинический исход, углубленный анализ данных.

1. Введение

В конце прошлого столетия, когда стала регистрироваться высокая частота возникновения неблагоприятных событий в процессе оказания медицинской помощи, в ряде развитых стран были проведены исследования, которые подтвердили, что вред здоровью, обусловленный не болезнью, а связанный с оказанием медицинской помощи, возникает у 10-15 % госпитализированных пациентов [1]-[4] В ряде стран были созданы подразделения, сообщающие об ошибках такого рода [5]-[7]. В России статистика врачебных ошибок не ведется- , хотя по неофициальным данным, ошибки медицинских работников уносят каждый год жизни около 50 тысяч человек [8].

Положительный опыт использования методов интеллектуального анализа данных (Big Data, Data Mining) во многих сферах (от финансов до управления сложными технологическими процессами), позволяет поставить вопрос о применимости этих методов и в решении задач прогнозирования риска возникновения неблагоприятных клинических исходов.

Целью настоящей работы является изложение первых результатов совместного проекта Национального медицинского исследовательского центра терапии и профилактической медицины (НМИЦТиПМ МЗ РФ) и механико-математического факультета МГУ имени М.В. Ломоносова, показывающего применимость методов анализа данных в решении этой важной проблемы. В рамках проекта решается задача прогнозирования

[1]работа выполнена при поддержке РФФИ грант № 19-29-01051 «Разработка алгоритмов принятия решений для управления рисками неблагоприятных клинических событий в высокотехнологичной медицинской организации на основе технологии data mining»

риска неблагоприятного клинического исхода на основе информации о пациенте, доступной в клинике НМИЦТиПМ МЗ РФ (медицинская информационная система «Медиалог»). Как и для всех проектов такого рода, заранее нельзя сказать, насколько успешным может быть этот подход к решению упомянутой выше задачи. Отчасти это связано со многими факторами неопределённости, для снятия которых необходимо провести серию экспериментальных работ и предварительно получить ответы на следующие вопросы:

- Что такое неблагоприятный клинический исход и можно ли его описать на языке накапливаемой в системе «Медиалог» информации о пациентах (признаки и их значения)? Изначально система «Медиалог» проектировалась для других целей и среди параметров системы признак неблагоприятного исхода отсутствует- поэтому ответ на этот вопрос далеко не очевиден.

- Достаточно ли данных (в смысле полноты представленной в системе информации о пациентах), чтобы такую задачу можно было решать или потребуется накапливать дополнительный набор признаков в описании (модели) пациента в системе?

- Достаточно ли представителен объем данных (в смысле количества пациентов) для возможности использования методов data mining (машинного обучения)?

- Достаточно ли данных (как количественных, так и качественных) накопленных в системе, чтобы к ним можно было достаточно эффективно применить методы data mining (машинного обучения)?

- Как оценивать качество прогноза? Какой уровень качества является приемлемым для клиницистов?

- Для использования клиницистами разрабатываемой системы прогнозирования, они должны доверять ей. Каким образом обеспечить такое доверие? Формальные критерии качества прогноза часто не убедительны для пользователей.

В ходе выполнения проекта были проведены несколько десятков встреч и совещаний между основными участниками проекта – сотрудниками клиники НМИЦТиПМ (будущими пользователями системы и носителями знаний о том, что «хорошо» и что «плохо»), сотрудниками ИТ подразделения НМИЦТиПМ (владельцами информации, понимающими все тонкости ее появления, накопления, изменения и хранения), сотрудниками механико-математического факультет МГУ имени М.В. Ломоносова (математиками, знающими как работают алгоритмы

анализа данных и как их настроить наилучшим образом для получения максимально хорошего результата). Первая часть обсуждений касалась понятия «неблагоприятный клинический исход» (основные участники – клиницисты и ИТ). Было сформулировано это понятие на языке признаков и их значений, имеющихся в системе. Вторая часть обсуждений касалась доступности и качества накапливаемой информации (основные участники – ИТ и математики). Было сделано несколько десятков выгрузок из системы (критерии уточнялись по результатам анализа предыдущей выгрузки), пока не были получены выгрузки с приемлемым уровнем пропуска значений признаков. Третья часть обсуждений касалась «доводки» полученной выгрузки до уровня, с которым могут работать алгоритмы анализа данных (основные участники – математики и клиницисты) с целью получения значимых для клиницистов результатов. Рассматривались разные варианты заполнения отсутствующих в выгрузке значений признаков (медианные значения/ усредненные/ дополнительный класс «-1» и пр.). Наряду с экспертными мнениями клиницистов (что «разумно», что «не разумно») использовались и формальные прогоны разными методами, позволяющие «нащупать» правильные алгоритмы заполнения недостающих данных. Четвертая часть обсуждений касалась выбора наилучших алгоритмов для полученного набора данных и их оптимизации (основные участники – математики и клиницисты). В частности, оказалось, что одним из важных критериев выбора класса алгоритмов является его интуитивная понятность для клиницистов. Для этого были разработаны методы выделения наиболее важных для прогноза признаков (это позволяло понять «разумность» прогноза), методы графического представления прогноза (это позволяло понять «логику» прогноза). Важность этого аспекта в начале проекта спрогнозировать было нельзя, поэтому пришлось делать много незапланированный работы. Тем не менее результаты работы этих алгоритмов были дополнительно проверены на менее понятной для клиницистов модели - нейронной сети. Результаты, полученные с помощью нейронной сети, показали хорошее качество предсказания риска развития неблагоприятных клинических исходов на отобранных ранее моделях.

В результате выполнения исследования были получены положительные ответы на сформулированные выше вопросы:

1) Задача может решаться на основе накапливаемой в ИТ системе НМИЦТиПМ информации;

2) Набор признаков (модель) достаточен, даже скорее избыточен, для применения методов анализа данных;

3) Качество информации находится в допустимых пределах, ее можно дополнить исходя из формальных соображений в случае необходимости;

4) Алгоритмы демонстрируют достаточно хорошее качество прогноза на имеющейся выборке;

5) Алгоритмы понятны врачам-клиницистам и они им доверяют.

Полученные результаты снимают потенциальные риски разработки прототипа системы, реализация которой составляет содержание следующих этапов проекта.

2. Обзор современного состояния исследований в области управления рисками неблагоприятных клинических событий в высокотехнологичной медицинской организации

Как уже упоминалось выше, возникновение неблагоприятных событий в процессе оказания медицинской помощи возникает у 10-15% госпитализированных пациентов. Один из главных путей решения этой важнейшей проблемы лежит в сфере информационных технологий, использование которых предоставляет возможность сделать оказание медицинской помощи более безопасным [9]-[11], а внедрение систем поддержки принятия клинических решений в управление качеством медицинской помощи дает возможность уменьшить врачебные ошибки, а также повысить безопасность хирургического лечения пациентов, оптимизировать процесс назначения лекарств и прогнозировать, а следовательно и уменьшать, риск неблагоприятных клинических событий [10]-[16]. Из немногочисленных примеров отечественных систем можно назвать клиническую информационную систему «ДОКА+» [17], ИС «Кардинет-онлайн» [18], которые содержат модуль, предупреждающий врачей о возможных ошибках с назначением лекарственных препаратов. Но системы, которые бы оказывали поддержку управленческим решениям, отсутствуют- во всяком случае нам не удалось найти информацию о таких системах.

Одной из нерешенных проблем использования информационных технологий в отечественном здравоохранении является слабое использование и управление медицинской информацией, которое затрудняет усилия по преобразованию ценности информации в ценность для отрасли [19], что приводит к росту расходов на медицинское обслуживание, к неоправданным затратам времени как для пациентов, так и для медицинских организаций. Все это обуславливает необходимость поиска эффективных

технологий, которые позволят медицинским организациям консолидировать организационные ресурсы для повышения безопасности и качества обслуживания пациентов, повышения организационной эффективности и, возможно, даже создания новых, более эффективных бизнес-моделей, основанных на использовании данных [20]-[22].

Многообещающим направлением в этой области является применение аналитики больших данных, которая включает в себя различные аналитические методы, такие как, например, описательная аналитика и интеллектуальная (прогнозная аналитика), которые идеально подходят для анализа большой части текстовых медицинских документов и других неструктурированных клинических данных (заметок, комментариев врача, медицинских изображений и пр.) [25]. Аналитика больших данных, разработанная на основе систем бизнес-аналитики позволит организациям здравоохранения анализировать огромный объем информации для принятия решений на основе фактических данных [23, 24].

Большинство моделей и методов, применяемых в настоящее время в различных отраслях медицины основаны на вероятностно-статистических моделях, которые сами по себе (как математические модели) неплохи, но часто дают плохие прогнозы, поскольку в приложениях начинают использовать неполные и недостоверные начальные данные, либо по малой выборке пытаются экстраполировать прогнозы на большие выборки и терпят неудачу. Кроме этого, сами модели используют методы, которые (в силу специфики современного образования) не понятны ни врачам-клиницистам, ни управленцам в медицинской области.

В настоящем исследовании предлагается использовать модели и методы обработки данных, возникших в технологиях интеллектуального (или углубленного) анализа данных (Big Data или Data Mining в англоязычной терминологии). Взаимодействие таких алгоритмов обработки данных может помочь в создании модели по выработке адекватных управленческих процессов с целью предотвращения неблагоприятных клинических событий, поскольку использует практически всю имеющуюся информацию, включая ту, которая не поддается обработке вероятностно-статистическими методами – в том числе слабоструктурированную, неполную, противоречивую и т.д.

Для отбора и предварительной обработки такой информации требуется наличие экспертов в предметной области и наличие самой информации, а также профессионалов - математиков по указанным выше областям. НМИЦТиПМ обладает достаточно репрезентативным корпусом информации (в том числе и в электронном виде) о пациентах, проходивших лечение в его клинике.

3. Используемый набор данных и предобработка данных

В современных высокотехнологических клиниках благодаря внедрению медицинских информационных систем (МИС) фиксируется и накапливается большой объем данных о каждом конкретном пациенте (демографические, клинические данные, результаты лабораторных и инструментальных методов диагностики, характер оперативного лечения и пр.), а также динамика происходящих с ним изменений.

В качестве данных для анализа были использованы 79 клинико- демографических и лабораторных параметров из МИС «Медиалог» НМИЦ-ТиПМ о 5062 пациентах, которым были выполнены высокотехнологичные эндоваскулярные (60%) и интервенционные аритмологические вмешательства (40%). В группу рентгенэндоваскулярных операций входили как диагностические (25% -коронарография, ангиография аорты и ее ветвей, ангиография брахиоцефальных артерий), так и лечебные процедуры (75% - ангиопластика и стентирование коронарных (77%), каротидных (7%) и периферических артерий (12,7%), транскатетерное протезирование аортального клапана (0,3%), ренальная денервация 3%). Аритмологические операции были представлены:

1) Внутрисердечными электрофизиологическими исследованиями;

2) Радиочастотными и криоаблациями аритмогенных очагов различных локализаций;

3) Имплантациями окклюзирующих устройств ушка левого предсердия;

4) Имплантациями электрокардиостимуляторов, кардиовертеров- дефибрилляторов, систем модулирующих сердечную сократимость, устройств для ресинхронизирующей терапии;

5) Имплантациями петлевых регистраторов ЭКГ

В среднем 15% операций проводились по экстренным показаниям.

По разным причинам исходные данные о пациентах содержали 58% пропусков (отсутствие записи), что считается довольно плохой оценкой для построения моделей машинного обучения. В связи с этим встал вопрос предобработки и очистки данных.

Целевой переменной являлся неблагоприятный клинический исход – adverse clinical event (ACE). Неблагоприятный клинический исход - это исход в периоперационном периоде, связанный с наступлением серьезного неблагоприятного сердечного-сосудистого и церебрального события

(смерть, инсульт, инфаркт миокарда и др.), а также клинически значимого кровотечения и/или специфического для каждого вида операций осложнения, приводящих к удлинению срока госпитализации и удорожанию стоимости лечения. В представленной выборке ACE принимал значение «1» в 84 случаях, значение «0» в 4978 случаях, соответственно. Так как целевая переменная принимала бинарные значения, то далее решалась задача классификации.

При анализе набора данных было обнаружено, что присутствуют некоторые признаки, которые можно не использовать в построении классификатора из-за того, что они либо неизвестны на момент поступления пациента в клинику, либо изначально не являются информативными. Такими признаками, например, являлись дата операции, дата выписки, количество койко-дней, id-пациента, город, должность и другие. Также было принято не использовать некоторые признаки, которые имеют достаточно большое количество пропусков. Такими признаками, например, являлись наличие плоского эпителия, солей в общем анализе мочи. В итоге был получен следующий набор данных в виде таблицы из 5062 строк (пациентов) и 66 столбцов (признаков). Процент пропусков, к сожалению, не изменился и составил 58%, поэтому была поставлена задача уменьшить процент пропусков в данных.

К этому вопросу нужно было подходить осторожно, поскольку размер исходных данных небольшой, а процент пропуска был довольно существенный. Существует несколько стандартных вариантов для решения:

- Отбрасывание записей. Это решение подходит только в том случае, если недостающие данные не являются информативными.
- Отбрасывание признаков. Такое решение может применяться только для неинформативных признаков.
- Внесение недостающих значений. Это решение подходит только в том случае, если в постановке задачи предусмотрено внесение искусственно созданных значений признаков пациента.
- Замена недостающих значений. Такое решение может применяться, когда возможны искусственные значения.

Для решения задачи нужны были данные только о реальных пациентах. Мы не могли просто отбросить все пропуски, так как имели небольшой объем данных, поэтому какой-то процент пропусков так или иначе должен быть. Было принято решение отбросить некоторое количество записей и признаков, учитывая особенности и размер исходных данных так, чтобы процент пропусков находился в промежутке 10-20%.

Эксперименты с набором данных показали, что для этого надо отбросить признаки, которые заполнены менее чем у 3000 пациентов, и отбросить пациентов, у которых заполнено менее чем 51 признак.

В итоге получился набор данных, содержащий сведения о 3146 пациентах и 23 признака, содержащих 16% пропусков. Задача достижения заданного процента пропусков с минимальным отбрасыванием данных нами не рассматривалась, поэтому описанная процедура, скорее всего, не является оптимальной. Результаты описанной процедуры предобработки данных в таблицах 1 и 2.

Таблица 1. Пропуски в исходном наборе данных

Номер	Признак	Количество пропусков
0	пол пациента	10
1	Натрий	2555
2	Кристаллы оксалатов	4324
3	Удельный вес	2993
4	АЧТВ	2927
5	Эритроциты (RBC)	349
6	Белок	3559
7	Аморфные фосфаты	4339
8	Гемоглобин (HGB)	341
9	Триглицериды	2708
10	Кристаллы трипельфосфатов	4325
11	Кислотность (pH)	2992
12	Фибриноген	2955
13	СОЭ(Скорость оседания эритроцитов)	1119
14	Цвет	419
15	Холестерин	2626
16	Лейкоциты (WBC)	340
17	Тромбоциты (PLT)	350
18	Калий	1934
19	Междунар. Нормализов. Отношение	2752
20	Гематокрит	349
21	Липопротеиды очень низкой плотности	4294
22	Креатинин	1741
23	Мочевая кислота	2883
24	ХС ЛПНП	3612
25	Глюкоза	2301
26	ХС ЛПВП	3594

Номер	Признак	Количество пропусков
27	Гликозилированный гемоглобин	4791
28	Ураты	4705
29	Гиалиновые цилиндры	3035
30	Протромбиновое время	2752
31	Неблагоприятное клиническое последствие	0
32	Диаметр аорты на уровне синусов	3487
33	Диаметр восходящей аорты	3422
34	Максимальное раскрытие створок аортного клапана	3483
35	Параметры систолического кровотока в выносящем тракте левого желудочка	3877
36	Параметры аортального кровотока: максимальная скорость	3434
37	Размер левого предсердия	3397
38	Объем левого предсердия	4217
39	Размер правого предсердия	3405
40	Объем правого предсердия	4463
41	Конечно-диастолический размер левого желудочка	3306
42	Конечно-систолический размер левого желудочка	3317
43	Толщина межжелудочковой перегородки в диастолу в выносящем тракте левого желудочка	3533
44	Толщина межжелудочковой перегородки	3408
45	Толщина задней стенки левого желудочка	3324
46	Масса миокарда левого желудочка	3597
47	Индекс массы миокарда левого желудочка	3660
48	Конечно-диастолический объем левого желудочка	3358
49	Конечно-систолический объем левого желудочка	3361
50	Ударный объем левого желудочка	3391
51	Минутный объем кровотока	3939
52	Глобальная сократимость левого желудочка фракция выброса	3301

Номер	Признак	Количество пропусков
53	Передне-заднее укорочение полости левого желудочка	3917
54	Максимальная скорость раннего диастолического наполнения	3398
55	Максимальная скорость кровотока во время предсердной систолы	3678
56	E/A	4500
57	Максимальный передне-задний размер правого желудочка	3330
58	Толщина передней стенки правого желудочка в диастолу	4086
59	Диаметр ствола легочной артерии	3461
60	Кровоток в стволе легочной артерии:	3511
61	Белок1	4320
62	ЧСС во время исследования	3434
63	PPT	3611
64	Год рождения	0
65	Сезон госпитализации	0

Таблица 2. Пропуски в исходном наборе данных

Номер	Признак	Количество пропусков
0	Пол пациента	7
1	Натрий	689
2	Удельный вес	1093
3	АЧТВ	1041
4	Эритроциты (RBC)	128
5	Гемоглобин (HGB)	126
6	Триглицериды	869
7	Кислотность (pH)	1093
8	Фибриноген	1066
9	СОЭ (Скорость оседания эритроцитов)	326
10	Цвет	138
11	Холестерин	789
12	Лейкоциты (WBC)	125
13	Тромбоциты (PLY)	129
14	Калий	413
15	Междунар.нормализов.отношение	956
16	Гематокрит	128

Номер	Признак	Количество пропусков
17	Креатинин	291
18	Мочевая кислота	989
19	Глюкоза	518
20	Протромбиновое время	956
21	Год рождения	0
22	Сезон госпитализации	0

Для заполнения пропусков в данных есть несколько стандартных вариантов:

- медианное значение
- среднее значение
- наиболее часто встречающееся значение
- искусственное значение (например, отрицательное).

После обсуждений с врачами НМИЦТиПМ было решено использовать медианное значение.

4. Решение задачи прогнозирования риска неблагоприятного клинического исхода

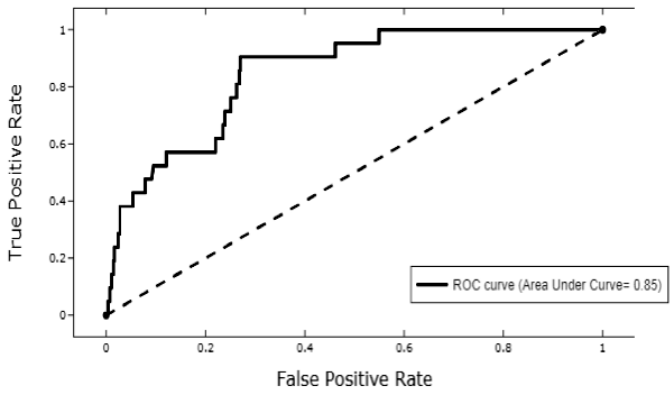
Для построения классификатора был выбран алгоритм построения дерева решений (Decision tree), так как требовалась хорошая интерпретируемость модели для медицинских специалистов, которые будут пользоваться классификатором. Именно в этом алгоритме, не прикладывая больших усилий, способен разобраться человек, далекий от математики. Для увеличения качества итогового дерева на исходных данных сначала использовался алгоритм случайного леса (Random Forest), заключающийся в использовании ансамбля решающих деревьев, каждое из которых само по себе даёт очень невысокое качество классификации, но за счёт их большого количества результат получается хорошим. Далее были выбраны 15 признаков с наибольшим вкладом в предсказание случайного леса, и итоговое дерево решений строилось на выборке, состоящей только из них.

Использовались меры качества алгоритма классификации F1-score и ROC AUC, так как они хорошо подходят для несбалансированных данных (значения целевой переменной относятся как 3079/67). Также для улучшения качества модели обучения использовалась кросс-валидация (cross-validation) - метод оценки аналитической модели и её поведения на независимых данных.

Во избежание переобучения алгоритмов Decision tree и Random forest были введены ограничения на их параметры:

- criterion - критерий, по которому происходит разбиение вершины: энтропия или Джини.
- max_depth - максимальная глубина дерева.
- min_samples_split - минимальное число элементов в вершине для разделения.
- min_samples_leaf - минимальное число элементов в листе.
- n_estimators - число деревьев в лесу. (только для Random forest)

Также был использован алгоритм случайного поиска лучших параметров классификатора по сетке - Random Search вместо перебора всевозможных вариантов - Grid Search по причине его трудоемкости.



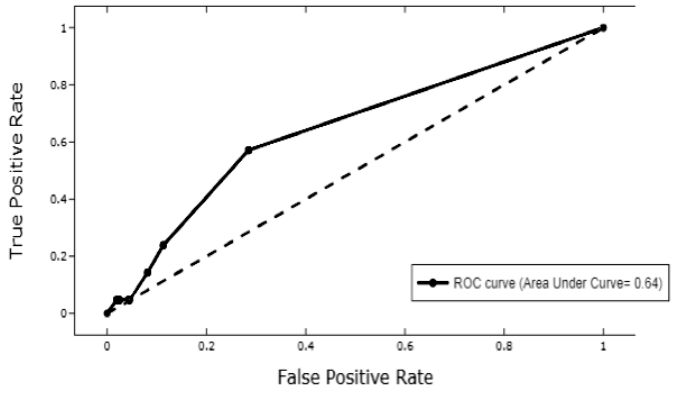
(a)

Для каждого алгоритма было проведено 50 запусков и выбран лучший классификатор по согласованным выше метрикам: случайный лес - F1-score = 0.97, ROC AUC = 0.85; дерево решений - F1-score = 0.96, ROC AUC = 0.64. Результаты представлены на рис. 1.

Отличие в значениях ROC AUC метрики на построенных классификаторах легко объясняется тем, что случайный лес изначально является более сильным алгоритмом, нежели дерево решений.

В рамках используемого подхода возможна оценка степени важности признака в решении задачи классификации. Таблица с первыми 15 наиболее важными признаками представлена в таблице 3.

Таблица 3. Список 15 наиболее важных признаков



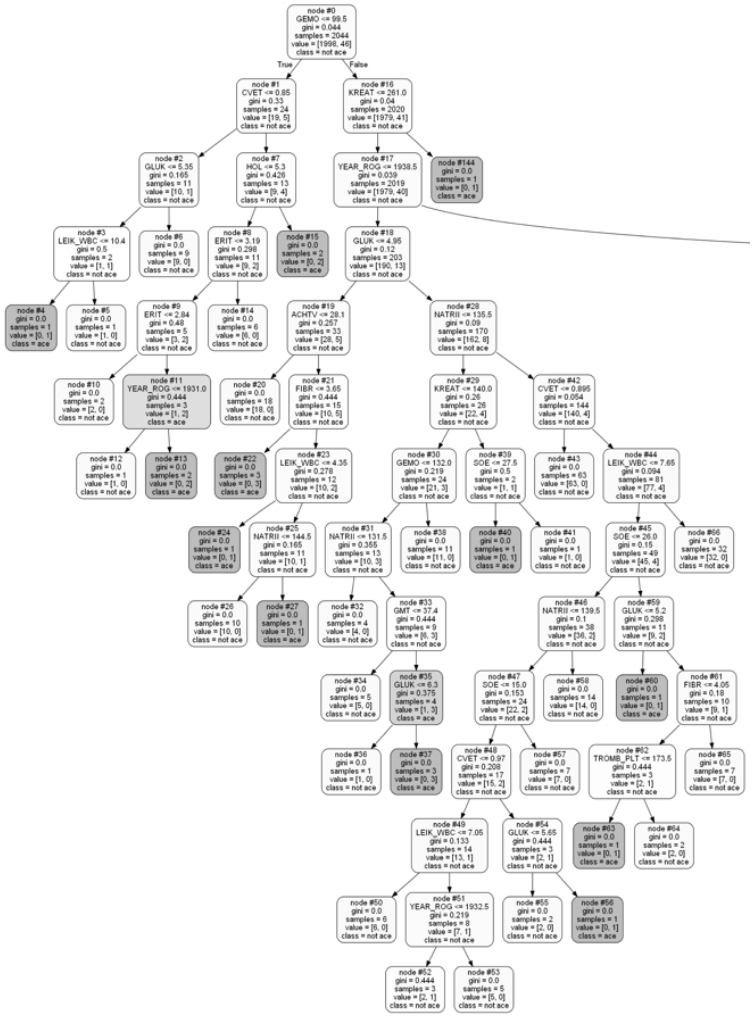
(б)

Рис. 1. ROC-кривая случайного леса (а) и дерева решений (б)

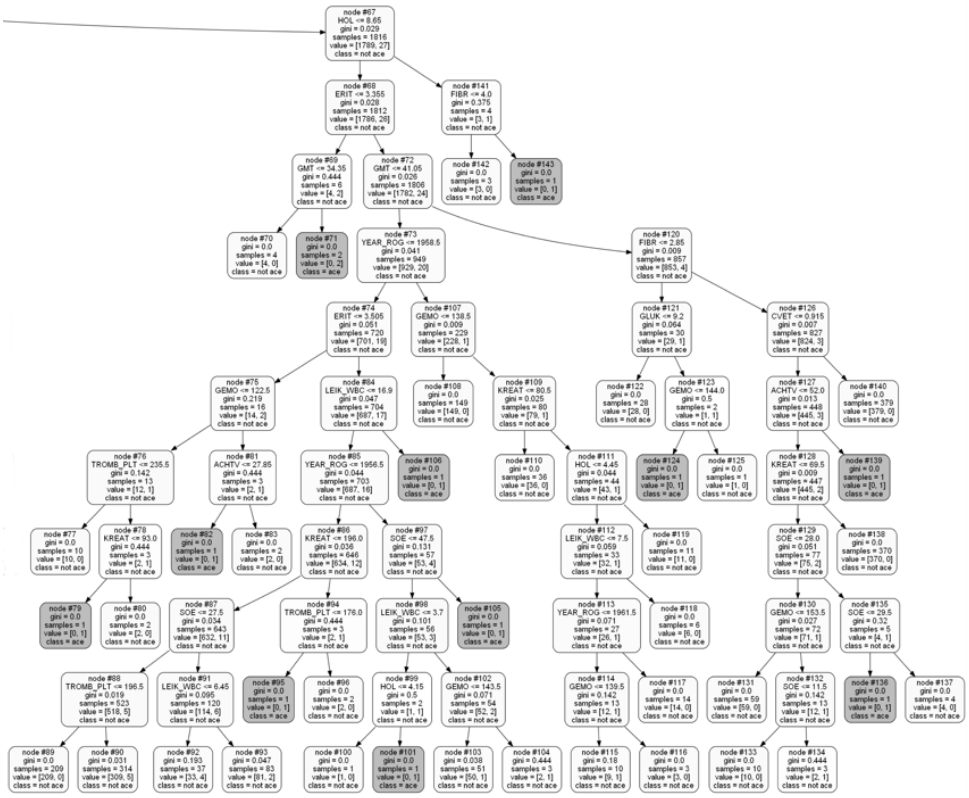
	index	field_name	feature_importances_rand	field_description
0	8.0	NATRII	0.120946	Натрий
1	3.0	SOE	0.118923	СОЭ(Скорость оседания эритроцитов)
2	10.0	TRIG	0.101272	Триглицериды
3	0.0	ERIT	0.088649	Эритроциты (RBC)
4	4.0	KREAT	0.083006	Креатинин
5	12.0	GLUK	0.071385	Глюкоза
6	13.0	LEIK_WBC	0.062446	Лейкоциты (WBC)
7	7.0	HOL	0.059461	Холестерин
8	11.0	ACHTV	0.056378	АЧТВ
9	1.0	GMT	0.046284	Гематокрит
10	6.0	TROMB_PLT	0.045132	Тромбоциты (PLT)
11	9.0	CVET	0.043179	Цвет
12	2.0	YEAR_ROG	0.041890	NaN
13	5.0	GEMO	0.040107	Гемоглобин (HGB)
14	14.0	KALII	0.020941	Калий

Отметим, что в целом этот набор признаков согласуется с экспертными представлениями. Некоторую дискуссию вызвала группа электролитов (натрий и калий), что может быть результатом ограниченности экспериментальной выборки и будет исследовано при разработке прототипа системы.

Структура итогового дерева решений представлена на рис. 2.


node #0
GEMO <= 99.5
gini = 0.044
samples = 2044
value = [1998, 46]
class = not ace
True
False
node #1
CVET <= 0.85
gini = 0.33
samples = 24
value = [19, 5]
class = not ace
node #16
KREAT <= 261.0
gini = 0.04
samples = 2020
value = [1979, 41]
class = not ace
node #2
GLUK <= 5.35
gini = 0.165
samples = 11
value = [10, 1]
class = not ace
node #7
HOL <= 5.3
gini = 0.426
samples = 13
value = [9, 4]
class = not ace
node #17
YEAR_ROG <= 1938.5
gini = 0.039
samples = 2019
value = [1979, 40]
class = not ace
node #144
gini = 0.0
samples = 1
value = [0, 1]
class = ace
node #3
LEIK_WBC <= 10.4
gini = 0.5
samples = 2
value = [1, 1]
class = not ace
node #6
gini = 0.0
samples = 9
value = [9, 0]
class = not ace
node #8
ERIT <= 3.19
gini = 0.298
samples = 11
value = [9, 2]
class = not ace
node #15
gini = 0.0
samples = 2
value = [0, 2]
class = ace
node #18
GLUK <= 4.95
gini = 0.12
samples = 203
value = [190, 13]
class = not ace
node #4
gini = 0.0
samples = 1
value = [0, 1]
class = ace
node #5
gini = 0.0
samples = 1
value = [1, 0]
class = not ace
node #9
ERIT <= 2.84
gini = 0.48
samples = 5
value = [3, 2]
class = not ace
node #14
gini = 0.0
samples = 6
value = [6, 0]
class = not ace
node #19
ACHTV <= 28.1
gini = 0.257
samples = 33
value = [28, 5]
class = not ace
node #28
NATRII <= 135.5
gini = 0.09
samples = 170
value = [162, 8]
class = not ace
node #10
gini = 0.0
samples = 2
value = [2, 0]
class = not ace
node #11
YEAR_ROG <= 1931.0
gini = 0.444
samples = 3
value = [1, 2]
class = ace
node #20
gini = 0.0
samples = 18
value = [18, 0]
class = not ace
node #21
FIBR <= 3.65
gini = 0.444
samples = 15
value = [10, 5]
class = not ace
node #29
KREAT <= 140.0
gini = 0.26
samples = 26
value = [22, 4]
class = not ace
node #42
CVET <= 0.895
gini = 0.054
samples = 144
value = [140, 4]
class = not ace
node #12
gini = 0.0
samples = 1
value = [1, 0]
class = not ace
node #13
gini = 0.0
samples = 2
value = [0, 2]
class = ace
node #22
gini = 0.0
samples = 3
value = [0, 3]
class = ace
node #23
LEIK_WBC <= 4.35
gini = 0.278
samples = 12
value = [10, 2]
class = not ace
node #30
GEMO <= 132.0
gini = 0.219
samples = 24
value = [21, 3]
class = not ace
node #39
SOE <= 27.5
gini = 0.5
samples = 2
value = [1, 1]
class = not ace
node #43
gini = 0.0
samples = 63
value = [63, 0]
class = not ace
node #44
LEIK_WBC <= 7.65
gini = 0.094
samples = 81
value = [77, 4]
class = not ace
node #24
gini = 0.0
samples = 1
value = [0, 1]
class = ace
node #25
NATRII <= 144.5
gini = 0.165
samples = 11
value = [10, 1]
class = not ace
node #31
NATRII <= 131.5
gini = 0.355
samples = 13
value = [10, 3]
class = not ace
node #38
gini = 0.0
samples = 11
value = [11, 0]
class = not ace
node #40
gini = 0.0
samples = 1
value = [0, 1]
class = ace
node #41
gini = 0.0
samples = 1
value = [1, 0]
class = not ace
node #45
SOE <= 28.0
gini = 0.15
samples = 49
value = [45, 4]
class = not ace
node #66
gini = 0.0
samples = 32
value = [32, 0]
class = not ace
node #26
gini = 0.0
samples = 10
value = [10, 0]
class = not ace
node #27
gini = 0.0
samples = 1
value = [0, 1]
class = ace
node #32
gini = 0.0
samples = 4
value = [4, 0]
class = not ace
node #33
GMT <= 37.4
gini = 0.444
samples = 9
value = [6, 3]
class = not ace
node #46
NATRII <= 139.5
gini = 0.1
samples = 38
value = [36, 2]
class = not ace
node #59
GLUK <= 5.2
gini = 0.298
samples = 11
value = [9, 2]
class = not ace
node #34
gini = 0.0
samples = 5
value = [5, 0]
class = not ace
node #35
GLUK <= 6.3
gini = 0.375
samples = 4
value = [1, 3]
class = ace
node #47
SOE <= 15.0
gini = 0.153
samples = 24
value = [22, 2]
class = not ace
node #58
gini = 0.0
samples = 14
value = [14, 0]
class = not ace
node #60
gini = 0.0
samples = 1
value = [0, 1]
class = ace
node #61
FIBR <= 4.05
gini = 0.18
samples = 10
value = [9, 1]
class = not ace
node #36
gini = 0.0
samples = 1
value = [1, 0]
class = not ace
node #37
gini = 0.0
samples = 3
value = [0, 3]
class = ace
node #48
CVET <= 0.97
gini = 0.208
samples = 17
value = [15, 2]
class = not ace
node #57
gini = 0.0
samples = 7
value = [7, 0]
class = not ace
node #62
TROMB_PLT <= 173.5
gini = 0.444
samples = 3
value = [2, 1]
class = not ace
node #65
gini = 0.0
samples = 7
value = [7, 0]
class = not ace
node #49
LEIK_WBC <= 7.05
gini = 0.133
samples = 14
value = [13, 1]
class = not ace
node #54
GLUK <= 5.65
gini = 0.444
samples = 3
value = [2, 1]
class = not ace
node #63
gini = 0.0
samples = 1
value = [0, 1]
class = ace
node #64
gini = 0.0
samples = 2
value = [2, 0]
class = not ace
node #50
gini = 0.0
samples = 6
value = [6, 0]
class = not ace
node #51
YEAR_ROG <= 1932.5
gini = 0.219
samples = 8
value = [7, 1]
class = not ace
node #55
gini = 0.0
samples = 2
value = [2, 0]
class = not ace
node #56
gini = 0.0
samples = 1
value = [0, 1]
class = ace
node #52
gini = 0.444
samples = 3
value = [2, 1]
class = not ace
node #53
gini = 0.0
samples = 5
value = [5, 0]
class = not ace

(a)



(б)

Рис. 2. Итоговое дерево решений: (а)-левая ветка, (б)-правая ветка

Кроме того, для проверки качества работы алгоритмов на исходных данных была создана и обучена нейронная сеть специального вида. Задача нейронной сети – выявление отклонений клинического состояния на начальном этапе их формирования.

Коротко опишем результаты использования этой модели. В качестве исходных данных использовались показатели 2959 пациентов по 23 признакам. Так как неблагоприятные события составляли 2% от доступных статистических данных, был разработан метод, используя который нейронная сеть была обучена в условиях отсутствия знаний о неблагоприятных событиях. Поскольку тип события может быть описан булевой логикой, то класс неблагоприятных событий может быть определен как отрицание благоприятного. То есть задача нейронной сети – выявить отклонение от благоприятного события относительно индивидуальных показателей пациента. В качестве исходного обучающего множества использовалось 2826 наблюдений благоприятных событий на основе 23 признаков, в качестве тестового множества использовалась совокупность из

66 неблагоприятных событий (2% от общего множества) и, сопоставимо, 66 благоприятных событий (выбраны случайным образом, в обучающем множестве не участвовали). Была разработана и обучена автоассоциативная нейронная сеть на основе алгоритма обратного распространение ошибки (градиентного спуска).

Топология сети: 23 нейрона входного слоя, 12 нейронов скрытого слоя, 23 нейрона выходного слоя (рис.3)

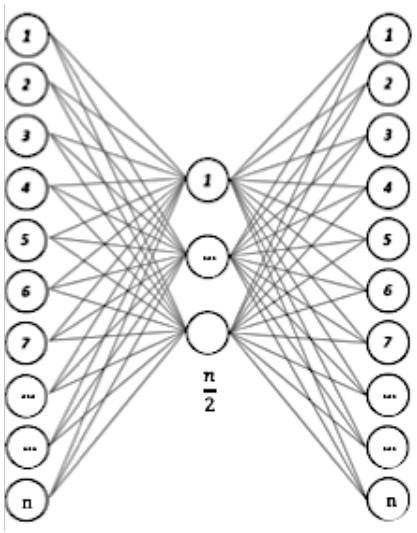


Рис. 3. Нейронная сеть

Результаты тестирования: на рассматриваемом тестовом множестве достигнуто 100% точности. Результат получен на основе 23 существенных признаков. При добавлении 12 несущественных признаков достигнута точность 96,96%, при этом 3% пришлось на ошибку второго рода, т.е. благоприятное событие принято за неблагоприятное.

Описание алгоритма: При обучении нейронная сеть в пространстве значительно меньшей размерности формирует комбинацию зависимостей, достаточных, чтобы на их основе интерпретировать в полной мере все 23 признака при обратном отображении. Нейронная сеть обучается до тех пор, пока не достигнет необходимой точности при обратном отображении. Скрытый слой может рассматриваться как нечеткое множество, содержащее 23 признака и функцию их принадлежности. Функция принадлежности формируется на основе весов связей между слоями. Таким образом зависимости фиксируются весами после обучения на максимально возможном множестве благоприятных событий. Далее при поступлении на вход неблагоприятного события возникают отклонения при

его обратном отображении. Событие определяется как неблагоприятное в случае возникновения отклонения хотя бы по одному из признаков.

Таблица 4. Список наиболее важных признаков

№	Признак	Важность
0	Пол пациента	0,238274
1	Удельный вес	0,220660
2	Натрий	0,161042
3	Эритроциты	0,147975
4	Гемоглобин (HGB)	0,138630
5	Гематокрит	0,137038
6	Рост пациента	0,129241
7	СОЭ (Скорость оседания эритроцитов)	0,114943
8	Холестерин	0,112352
9	Глюкоза	0,111754
10	Цвет	0,109221
11	Тромбоциты (PLT)	0,095458
12	Триглицериды	0,094066
13	Кислотность	0,090038
14	Протромбиновое время	0,088945
15	Мочевая кислота	0,083856
16	Междунар. нормализов. отношение	0,083054
17	Креатинин	0,082132
18	Лейкоциты (WBC)	0,071010
19	Вес пациента	0,065011
20	Калий	0,064191
21	Фибриноген	0,054971
22	АЧТВ	0,054791

Степень значимости признака определяется на основе функций принадлежности признаков к нечеткому подмножеству, то есть на основе весов между первым и скрытым слоями. После обучения нейронной сети на множестве благоприятных событий может быть составлен рейтинг значимости признаков. Степени значимости признаков позволяют объяснить принятое нейронной сетью решение, поскольку отклонение по каждому из неблагоприятных событий фиксируется на основе конкретного признака.

Полученные результаты демонстрируют высокую эффективность разработанной нейронной сети при применении на медицинских данных подобного рода и подтверждают применимость подобной архитектуры

алгоритма для решения рассматриваемой задачи. Нейронная сеть была разработка на языке Python без использования специальных пакетов. Для операций с многомерными массивами использовались функции библиотеки NumPy, для предобработки и нормализации данных использовались функции библиотеки Pandas.

5. Заключение

Итоги проведенного исследования показали, что методы машинного обучения можно применять для анализа медицинских данных. Показано, что можно предсказывать неблагоприятный клинический исход пациента при поступлении в клинику, причем с хорошей точностью.

В ходе исследования были выявлены показатели, которые имели наибольшую прогностическую ценность (таблица 1.) Эндоваскулярные и аритмологические операции, несмотря на то, что имеют схожие черты и относятся к разновидностям малоинвазивной интервенционной радиологии, существенно отличаются друг от друга характером проводимого лечения и особенностями периоперационных осложнений. В литературных источниках доступна информация о предикторах неблагоприятного клинического исхода только для определенного типа операций. Однако такие клинико-демографические параметры, как пожилой возраст пациента, наличие сопутствующих анемии (гемоглобин, гематокрит), воспалительного процесса (лейкоциты, СОЭ), нарушений в свертывающей системы крови (тромбоциты), метаболических (холестерин, триглицериды; глюкоза) и электролитных нарушений отражают неблагоприятное клиническое состояние пациента и ассоциированы с повышенным риском развития операционных осложнений, что нашло подтверждение в многочисленных клинических исследованиях и регистрах.

Полученные результаты могут быть использованы в качестве рекомендации для медицинских работников при первичном осмотре поступивших пациентов, а также позволяют разработать систему прогнозирования риска неблагоприятного клинического исхода по технологии скоринговых систем. Получен вклад наиболее важных переменных в предсказание риска неблагоприятного клинического исхода. Имея значения этих признаков, медицинский работник может самостоятельно оценить риск неблагоприятного клинического исхода.

Список литературы

[1] Brennan T.A. et al., “Incidence of adverse events and negligence in hospitalized patients: results of the Harvard Medical Practice Study I”, *N. Engl. J. Med. Mass Medical Soc*, **324**:6 (1991), 370–376.

[2] Vincent C., Neale G., Woloshynowych M., “Adverse events in British hospitals: preliminary retrospective record review”, *Bmj. British Medical Journal Publishing Group*, **322**:7285 (2001), 517–519.

[3] Wilson R.M. et al, “The quality in Australian health care study”, *Med. J. Aust. Sydney, Australia: Australian Medical Association*, **163**:9 (1914-, 1995), 458–471.

[4] Baker G.R. et al, “The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada”, *Can. Med. Assoc. J. Can Med Assoc*, **170**:11 (2004), 1678–1686.

[5] Runciman W.B., “246–251”, *BMJ Qual. Saf. BMJ Publishing Group Ltd*, **11**:3 (2002).

[6] Kohn L.T., Corrigan J.M., Donaldson M.S., “Error reporting systems”, *National Academies Press (US)*, 2000.

[7] Mayor S., “English NHS to set up new reporting system for errors”, *BMJ Br. Med. J. BMJ Publishing Group LTD*, **320**:7251 (2000), 1689.

[8] Лудупова Е.Ю., “Врачебные ошибки. Литературный обзор”, *Вестник Росздравнадзора. Федеральное государственное бюджетное учреждение"Информационно-методический*, 2016, № 2, 6–15.

[9] Kass B.L., “Reducing and preventing adverse drug events to decrease hospital costs”, *Res. action*, **1** (2001), 1–20.

[10] Bates D.W., Gawande A.A., “Improving safety with information technology”, *N. Engl. J. Med. Mass Medical Soc*, **348**:25 (2003), 2526–2534.

[11] Parente S.T., McCullough J.S., “Health information technology and patient safety: evidence from panel data”, *Health Aff. Project HOPE-The People-to-People Health Foundation, Inc.*, **28**:2 (2009), 357–360.

[12] Jao C.S., Hier D.B., “Clinical decision support systems: An effective pathway to reduce medical errors and improve patient safety”, *Decision Support Systems. InTech*, 2010.

[13] Bates D.W. et al., “Reducing the frequency of errors in medicine using information technology”, *J. Am. Med. Informatics Assoc. BMJ Group BMA House, Tavistock Square, London, WC1H 9JR*, **8**:4 (2001), 299–308.

[14] Berner E.S., “Clinical decision support systems”, *Springer*, **233** (2007).

[15] Chaudhry B., “Computerized clinical decision support: will it transform healthcare?”, *Springer*, 2008.

[16] Weaver C.A. et al., “Healthcare information management systems”, *Cham Springer Int. Publ. Springer*, 2016.

[17] Шульман Е.И. et al., “Технические решения, свойства и возможности клинической информационной системы ДОКА”, *Системы*, 2009, 88.

[18] Атьков О.Ю. et al., “Система поддержки принятия врачебных решений”, *Врач и информационные технологии. Общество с ограниченной ответственностью Издательский дом «Менеджер*, 2013, № 6.

[19] Goodman J., Gorman L., Herrick D., “Health Information Technology: Benefits and Problems”, *Natl. Cent. Policy Anal. Washingt*, 2010.

[20] Agarwal R. et al., “Research commentary—The digital transformation of healthcare: Current status and the road ahead”, *Inf. Syst. Res. INFORMS*, **21**:4 (2010), 796–809.

[21] Goh J.M., Gao G., Agarwal R., “Evolving work routines: Adaptive routinization of information technology in healthcare”, *Inf. Syst. Res. INFORMS*, **22**:3 (2011), 565–585.

[22] Ker J.-I. et al., "Deploying lean in healthcare: Evaluating information technology effectiveness in US hospital pharmacies", *Int. J. Inf. Manage. Elsevier*, **34**:4 (2014), 556–560.

[23] Ikehara S. et al., "Alcohol consumption and mortality from stroke and coronary heart disease among Japanese men and women: the Japan collaborative cohort study", *Stroke. Am Heart Assoc*, **39**:11 (2008), 2936–2942.

[24] Raghupathi W., Raghupathi V., "Big data analytics in healthcare: promise and potential", *Heal. Inf. Sci. Syst. BioMed Central*, **2**:1 (2014), 3.

[25] Groves P. et al., "The "big data" revolution in healthcare", *Accelerating value and innovation. McKinsey Company*, 2016.

The adverse clinical outcome risk assessment by in-depth data analysis methods

Gornyi B.E., Ryjov A.P., Strogalov A.S., Zhuravlev A.D., Khusaenov A.A., Shergin I.A., Feshchenko D.A., Abdullaev A.M., Kontsevaya A.V.

The adverse events in the medical care providing process occurs in 10-15% of hospitalized patients. Even a few percent reducing of such events will save thousands of lives. One of the ways to solve this crucial problem is the usage of intelligent information technologies that allow the predicting of an unfavorable clinical outcome risk in patients. The paper presents the study results carried out jointly by the scientist of the National Research Center for Therapy and Preventive Medicine of the Ministry of Health of the Russian Federation and the the scientist of Faculty of Mechanics and Mathematics of the Lomonosov Moscow State University, showing the applicability of data analysis methods in this important problem solving.

Keywords preventive medicine, adverse clinical outcome, in-depth data analysis.

References

[1] Brennan T.A. et al., "Incidence of adverse events and negligence in hospitalized patients: results of the Harvard Medical Practice Study I", *N. Engl. J. Med. Mass Medical Soc*, **324**:6 (1991), 370–376.

[2] Vincent C., Neale G., Woloshynowych M., "Adverse events in British hospitals: preliminary retrospective record review", *Bmj. British Medical Journal Publishing Group*, **322**:7285 (2001), 517–519.

[3] Wilson R.M. et al, "The quality in Australian health care study", *Med. J. Aust. Sydney, Australia: Australian Medical Association*, **163**:9 (1914-, 1995), 458–471.

[4] Baker G.R. et al, "The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada", *Can. Med. Assoc. J. Can Med Assoc*, **170**:11 (2004), 1678–1686.

[5] Runciman W.B., "246–251", *BMJ Qual. Saf. BMJ Publishing Group Ltd*, **11**:3 (2002).

[6] Kohn L.T., Corrigan J.M., Donaldson M.S., "Error reporting systems", *National Academies Press (US)*, 2000.

[7] Mayor S., "English NHS to set up new reporting system for errors", *BMJ Br. Med. J. BMJ Publishing Group LTD*, **320**:7251 (2000), 1689.

[8] Ludupova E.Y., "Medical errors. Literature review", *VESTNIK ROSZDRAVNADZORA. Federal State Budgetary Institution "Information and Methodological Center for Expert Evaluation, Recording and Analysis of Circulation of Medical Products"*, 2016, № 2, 6–15 (In Russian).

[9] Kass B.L., "Reducing and preventing adverse drug events to decrease hospital costs", *Res. action*, **1** (2001), 1–20.

[10] Bates D.W., Gawande A.A., "Improving safety with information technology", *N. Engl. J. Med. Mass Medical Soc*, **348**:25 (2003), 2526–2534.

[11] Parente S.T., McCullough J.S., "Health information technology and patient safety: evidence from panel data", *Health Aff. Project HOPE-The People-to-People Health Foundation, Inc.*, **28**:2 (2009), 357–360.

[12] Jao C.S., Hier D.B., "Clinical decision support systems: An effective pathway to reduce medical errors and improve patient safety", *Decision Support Systems. InTech*, 2010.

[13] Bates D.W. et al., "Reducing the frequency of errors in medicine using information technology", *J. Am. Med. Informatics Assoc. BMJ Group BMA House, Tavistock Square, London, WC1H 9JR*, **8**:4 (2001), 299–308.

[14] Berner E.S., "Clinical decision support systems", *Springer*, **233** (2007).

[15] Chaudhry B., "Computerized clinical decision support: will it transform healthcare?", *Springer*, 2008.

[16] Weaver C.A. et al., "Healthcare information management systems", *Cham Springer Int. Publ. Springer*, 2016.

[17] Shulman E.I. et al., "Technical solutions, properties and capabilities of the DOCA clinical information system", *Systems*, 2009, 88 (In Russian).

[18] Atkov O.Y. et al., "Clinical decision support system", *Information technologies for the Physician. Manager of Health Care*, 2013, № 6 (In Russian).

[19] Goodman J., Gorman L., Herrick D., "Health Information Technology: Benefits and Problems", *Natl. Cent. Policy Anal. Washingt*, 2010.

[20] Agarwal R. et al., "Research commentary—The digital transformation of healthcare: Current status and the road ahead", *Inf. Syst. Res. INFORMS*, **21**:4 (2010), 796–809.

[21] Goh J.M., Gao G., Agarwal R., "Evolving work routines: Adaptive routinization of information technology in healthcare", *Inf. Syst. Res. INFORMS*, **22**:3 (2011), 565–585.

[22] Ker J.-I. et al., "Deploying lean in healthcare: Evaluating information technology effectiveness in US hospital pharmacies", *Int. J. Inf. Manage. Elsevier*, **34**:4 (2014), 556–560.

[23] Ikehara S. et al., "Alcohol consumption and mortality from stroke and coronary heart disease among Japanese men and women: the Japan collaborative cohort study", *Stroke. Am Heart Assoc*, **39**:11 (2008), 2936–2942.

[24] Raghupathi W., Raghupathi V., "Big data analytics in healthcare: promise and potential", *Heal. Inf. Sci. Syst. BioMed Central*, **2**:1 (2014), 3.

[25] Groves P. et al., "The "big data" revolution in healthcare", *Accelerating value and innovation. McKinsey Company*, 2016.